The Cambridge Handbook of RESPONSIBLE ARTIFICIAL INTELLIGENCE

Interdisciplinary Perspectives

EDITED BY

Silja Voeneky, Philipp Kellmeyer, Oliver Mueller and Wolfram Burgard

CAMBRIDGE

THE CAMBRIDGE HANDBOOK OF RESPONSIBLE ARTIFICIAL INTELLIGENCE

In the past decade, artificial intelligence (AI) has become a disruptive force around the world, offering enormous potential for innovation but also creating hazards and risks for individuals and the societies in which they live. This volume addresses the most pressing philosophical, ethical, legal, and societal challenges posed by AI. Contributors from different disciplines and sectors explore the foundational and normative aspects of responsible AI and provide a basis for a transdisciplinary approach to responsible AI. This work, which is designed to foster future discussions to develop proportional approaches to AI governance, will enable scholars, scientists, and other actors to identify normative frameworks for AI to allow societies, states, and the international community to unlock the potential for responsible innovation in this critical field. This book is also available as Open Access on Cambridge Core.

SILJA VOENEKY is a leading scholar in the field of the interdependence of ethics and public international law, and the governance of emerging technologies. She is Professor of Public International Law, and Comparative Law at the University of Freiburg and was a Fellow at Harvard Law School. She previously served as a member of the German Ethics Council. Since 2001, Voeneky has been, inter alia, a legal advisor to the German Federal Foreign Office and the German Federal Ministry of Environment.

PHILIPP KELLMEYER is a neurologist and neuroscientist at the Medical Center – University of Freiburg. In his neuroscientific work, he investigates language processing in the brain and the clinical application of brain–computer interfaces, virtual reality, and other digital technologies. Through his contributions to neuroethics, he has become an international expert on ethical aspects of neuroscience, neurotechnology, and AI.

OLIVER MUELLER is Professor of Philosophy with a focus on technology and on contemporary philosophy, University of Freiburg, and a Senior Fellow at the 2018–2021 FRIAS Saltus Research Group 'Responsible AI'.

WOLFRAM BURGARD is a leading researcher in Robotics and Artificial Intelligence. He co-authored more than 400 publications including the famous book 'Probabilistic Robotics'. He received numerous awards including the Gottfried Wilhelm Leibniz Prize, the most prestigious German research award, and is Fellow of several societies.

The Cambridge Handbook of Responsible Artificial Intelligence

INTERDISCIPLINARY PERSPECTIVES

Edited by

SILJA VOENEKY

University of Freiburg

PHILIPP KELLMEYER

Medical Center – University of Freiburg

OLIVER MUELLER

University of Freiburg

WOLFRAM BURGARD

University of Technology Nuremberg

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009207867

DOI: 10.1017/9781009207898

First published 2022

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
NAMES: Voeneky, Silja, 1969– editor. | Kellmeyer, Philipp, 1979– editor. | Mueller, Oliver, 1972– editor. | Burgard, Wolfram, editor.
TITLE: The Cambridge handbook of responsible artificial intelligence : interdisciplinary perspectives / edited by Silja Voeneky, University of Freiburg; Philipp Kellmeyer, Medical Center – University of Freiburg; Oliver Mueller, University of Freiburg; Wolfram Burgard, University of Technology Nuremberg.
DESCRIPTION: Cambridge, United Kingdom; New York, NY: Cambridge University Press, 2022. | Series: Cambridge law handbooks | Includes index.
IDENTIFIERS: LCCN 2022022867 (print) | LCCN 2022022868 (ebook) | ISBN 9781009207867 (hardback) | ISBN 9781009207898 (epub)
SUBJECTS: LCSH: Artificial intelligence–Law and legislation. | Artificial intelligence–Social aspects. | BISAC: LAW / General
CLASSIFICATION: LCC K564.C6 C3595 2022 (print) | LCC K564.C6 (ebook) | DDC 343.09/99–dc23/eng/20220630
LC record available at https://lccn.loc.gov/2022022867
LC ebook record available at https://lccn.loc.gov/2022022868

ISBN 978-1-009-20786-7 Hardback

Contents

Figures

Contributors

Ebrahim Afsah is an associate professor for public international law at the University of Copenhagen (Denmark); he was a professor of Islamic law at the University of Vienna (Austria), and a visiting professor of international relations at the University of Brest (France). He was trained at the School of Oriental and African Studies, London, Trinity College Dublin, the Kennedy School of Government at Harvard University, and the Max Planck Institute for Comparative Public and International Law in Heidelberg (Germany). In recent years, *Ebrahim Afsah* has won fellowships to the European University Institute in Florence (Fernand Braudel), the Norwegian Academy of Science (Nordic Civil Wars), Harvard Law School (Islamic Legal Studies Program), and the National University of Singapore (Centre for Asian Legal Studies).

Wolfram Burgard is a professor of computer science at the University of Technology Nuremberg (UTN), and was a professor of computer science at the University of Freiburg (Germany), where he headed the research lab for Autonomous Intelligent Systems. Besides, he was a Senior Fellow at the 2018–2021 FRIAS Saltus Research Group 'Responsible AI - Emerging ethical, legal, philosophical and social aspects of the interaction between humans and autonomous intelligent systems.' His areas of interest lie in AI and mobile robots. His research focuses on the development of robust and adaptive techniques for state estimation and control. Over the past years, his group and he have developed a series of innovative probabilistic techniques for robot navigation and control. They cover different aspects such as localization, map-building, SLAM, path-planning, and exploration. *Wolfram Burgard* has published more than 400 papers and articles in robotic and AI conferences and journals. He is coauthor of two books, *Principles of Robot Motion: Theory, Algorithms, and Implementations* and *Probabilistic Robotics*. He is a Fellow of the European Association for Artificial Intelligence (EurAI), the Association for the Advancement of Artificial Intelligence (AAAI), and the Institute of Electrical and Electronics Engineers (IEEE). He is, furthermore, a member of the German Academy of Sciences Leopoldina as well as of the Heidelberg Academy of Sciences and Humanities.

Thomas Burri is a professor of international law and European law at the University of St. Gallen (Switzerland). His research has been published in numerous international outlets. He has published three books, *The International Court of Justice and Decolonisation* (ed. with Jamie Trinidad, Cambridge University Press 2021), *The Greatest Possible Freedom* (Nomos 2015), and *Models of Autonomy*? (Schulthess 2010). His research covers traditional international law and EU law, but *Thomas Burri* has also investigated AI and autonomous systems with empirical methods for more than a decade.

Christoph Durt is currently a postdoctoral researcher at the Freiburg Institute of Advanced Studies (FRIAS, Germany). The core topic of his research is the interrelation of computational technology with human experience and thought. He combines systematic philosophy with interdisciplinary collaboration and conceptual investigation of the developments that underlie the digital age. Before coming to Freiburg, he was the scientific head of several interdisciplinary projects on consciousness, self, and AI, funded by the European Union's Horizon 2020 research and innovation program and the Volkswagen Foundation. He has taught on a wide range of topics and authors from Ancient Philosophy to the present at Munich University, the University of California, Santa Cruz, the University of California, Berkeley, and the University of Vienna. He has received several research and teaching awards, including one for the essay 'The Computation of Bodily, Embodied, and Virtual Reality' and another for the essay 'How the Digitization of Our World Changes Our Orientation' (links and more on: www.durt.de).

Boris Essmann studied philosophy, cognitive science, and anthropology at Freiburg University (Germany) and is currently finishing his PhD project. He has worked on several research projects in the field of neurophilosophy and neuroethics, including the Bernstein Focus Neurotechnology Freiburg/Tübingen (2010–2012) and a neurophilosophical Junior Research Group in BrainLinks-BrainTools at the University of Freiburg (2013–2018). He currently works at the EU project FUTUREBODY (2018–2020).

Johanne Giesecke is an articled clerk in the Dresden Higher Regional Court district (Germany). She studied law at the University of Heidelberg (Germany) and at the Université de Montpellier (France), the latter with an Erasmus scholarship. She completed her first state exam in 2021 in Heidelberg. From 2018 until 2021, she was a student assistant in the research group of Prof. Dr. *Fruzsina Molnár-Gábor* in various projects in the field of medical and health law at the Heidelberg Academy of Sciences and Humanities.

Jan von Hein is a Director at the Institute for Comparative and Private International Law at the University of Freiburg (Germany). He is the chairman of the Second Commission of the German Council for Private International Law, a member of the Board of the International Law Association's German branch and an associate member of the International Academy of Comparative Law. He is the author of numerous books and articles on private international and comparative law, which have been honoured by the Max Planck Society and the German Stock Corporation Institute.

Wilfried Hinsch is a full university professor and holds the chair for Practical Philosophy at the University of Cologne (Germany). His most recent book publications are *Die gerechte Gesellschaft* (Stuttgart 2016), *Die Moral des Krieges* (München 2018) and *Öffentliche Vernunft? Die Wissenschaft in der Demokratie* (co-edited with Daniel Eggers, Berlin/Boston 2019). Currently, he is working on a book manuscript with the title *Legitimacy Beyond Procedural Justice*.

Philipp Kellmeyer, is a neurologist at the Medical Center – University of Freiburg (Germany), where he heads the Human-Technology Interaction Lab at the Department of Neurosurgery. He was a Senior Fellow at the 2018–2021 FRIAS Saltus Research Group 'Responsible AI – Emerging ethical, legal, philosophical, and social aspects of the interaction between humans and autonomous intelligent systems'. *Philipp Kellmeyer* studied human medicine in Heidelberg and Zurich and received a Master of Philosophy from the University of Cambridge (UK). As a neuroscientist, he works in the fields of neuroimaging and translational neurotechnology, in particular the clinical application of AI-based brain–computer interfaces. He is a scientific

member of the BrainLinks-BrainTools centre at the University of Freiburg. In his neuroethical research, he works on ethical, legal, social, and political challenges of neurotechnologies, big data, and AI in medicine and research. He is also a research affiliate of the Institute for Biomedical Ethics and History of Medicine at the University of Zurich, where he also teaches biomedical ethics.

Haksoo Ko is a professor of law at Seoul National University School of Law in Seoul (Korea). He primarily teaches areas in law and economics as well as in data privacy and AI law. He regularly sits on various advisory committees for the Korean government and other public and private institutions. Prior to joining academia, he practiced law with law firms in New York and in Seoul. He currently serves as President of Asian Law and Economics Association; President of Korean Association for AI and Law; Co-director of SNU AI Policy Initiative; and Associate Director of SNU AI Institute. He had visitor appointments at UC Berkeley, University of Hamburg, Vrije Universiteit Brussel, University of Freiburg, and National University of Singapore. He holds a B.A. in Economics from Seoul National University and received both J.D. and Ph.D. (Economics) degrees from Columbia University in New York, USA.

Christoph Krönke is a full university professor of public law at Vienna University of Economics and Business (WU, Austria). His research interests include the law of digitalisation, in particular the law of the digital economy. He studied law at Ruprecht Karl University in Heidelberg (2003–2005) and Ludwig Maximilian University Munich (2005–2009). He received his doctorate in law in 2013 with a dissertation on European and administrative law. In his habilitation thesis of 2020, he developed the foundations of the 'Regulation of the Digital Economy', with a focus on the regulation of digital platforms and networks as well as intelligent systems.

Alex Leveringhaus is a lecturer on political theory at the University of Surrey (UK). Furthermore, he is the Co-director of the Centre for International Intervention. Beforehand, *Alex Leveringhaus* was a Leverhulme Early Career Research Fellow in the Centre for Political Theory at the University of Manchester (UK). Prior to this, he was a postdoctoral research Fellow at the Oxford Institute for Ethics, Law and Armed Conflict (UK). Generally, his research focuses on the ethical and political repercussions of the widespread introduction and use of AI. In 2016 he published his monograph, *Ethics and Autonomous Weapons*.

Dustin A. Lewis is the Research Director for the Harvard Law School Program on International Law and Armed Conflict (HLS PILAC) (USA). With a focus on public international law sources and methods, he leads research into several wide-ranging contemporary challenges concerning armed conflict. Among his recent areas of focus, *Dustin Lewis* led the HLS PILAC project on 'International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns'. Alongside a team of faculty and research assistants, he explored how international law governs the development and use in war of AI and certain other advanced algorithmic and data-reliant socio-technical systems. As part of that project and its forerunner, he presented the program's research on AI, international law, and armed conflicts in Beijing, Geneva, Moscow, New Delhi, Oxford, Shanghai, Stockholm, and Washington, D.C. He is an honours graduate of Harvard College (A.B.) and Utrecht University School of Law (LL.M.).

Jan Lieder is a full professor, holds the chair for Civil Law, Commercial Law and Business Law and is the Director of the Department of Business Law at the Institute for Business Law, Labor and Social Law at the University of Freiburg (Germany). Moreover, he is a judge at the Higher Regional Court Schleswig. In 2017, *Jan Lieder* was a Stanford Law Senior Visiting Scholar. He

attained an LLM from Harvard Law School with a concentration in corporate law. His main areas of research include German, European, and international commercial law, corporate law, and business law. In these fields of law, he is an author of more than 200 publications and articles.

Yong Lim is an associate professor at Seoul National University, School of Law (Korea). He is also the Co-director of the SNU AI Policy Initiative at SNU's Center for Law and Economics. He has graduated from Seoul National University, College of Law, and obtained his S.J.D. at Harvard Law School. Prior to joining academia, *Yong Lim* practiced law at Korea's leading law firm specializing in antitrust and competition law. His areas of specialty also include consumer protection, information technology law, and privacy and data protection law.

Yun Liu is a Postdoc at the Tsinghua University School of Law and Assistant Director of the Institute for Studies on Artificial Intelligence and Law, Tsinghua University (China). In 2017, he received his Ph.D. at the China University of Political Science and Law. From August 2016 to August 2017, he was a visiting scholar at the Ohio State University. His research interests include computational law, standardization law, and civil law. He has worked as an expert in the legislation on technical standards and data governance. Since 2018, he has served in a LegalTech research program, organizing computer scientists and jurists to develop intelligence tools for judicial trials on civil and commercial cases in Tsinghua University. He has published more than twenty papers in Chinese on AI governance, data governance, technical standards, and civil law.

Thomas Metzinger was a full professor of theoretical philosophy at the Johannes Gutenberg-Universität Mainz until 2019 (Germany); from 2019 to 2021 he was awarded a Senior Forschungsprofessur by the Ministry of Science, Education and Culture. He is past president of the German Cognitive Science Society (2005–2007) and of the Association for the Scientific Study of Consciousness (2009–2011). As of 2011, he is an Adjunct Fellow at the Frankfurt Institute for Advanced Studies; from 2008 to 2009 he served as a Fellow at the Berlin Institute for Advanced Study; from 2014 to 2019, he was a Fellow at the Gutenberg Research College. From 2018 to 2020 *Thomas Metzinger* worked as a member of the European Commission's High-Level Expert Group on AI.

Catrin Misselhorn holds a chair for philosophy at the University of Göttingen (Germany). From 2012 until 2019 she was chair for the philosophy of science and technology at the University of Stuttgart. Prior to that she was visiting professor at the Humboldt University in Berlin, the University of Zurich, and the University of Tübingen. In 2003 she received her Ph.D. at the University of Tübingen and in 2010 she finished her habilitation. In the years 2007–2008 she was Feodor Lynen research Fellow at the Center of Affective Sciences in Geneva, at the Collège de France, and the Institut Jean Nicod for cognitive science in Paris. Her main research areas are philosophical problems in AI, robot and machine ethics, and human–machine interaction. She is leading a number of third party funded projects on the ethical assessment of assistive systems in different areas, for instance, in care, at the workplace, and in education.

Fruzsina Molnár-Gábor is a professor at the Faculty of Law, University of Heidelberg (Germany) and a member of the Akademie Kolleg of the Heidelberg Academy of Sciences and Humanities. Her research focuses on the regulation of biomedicine and biotechnology, including the fields of data law, medical law, international law, and the law of the European Union. She is a member of the Junge Akademie of the Berlin-Brandenburg Academy of

Sciences and Humanities, the Leopoldina, and the European Group on Ethics in Science and New Technologies (EGE). She has received the Manfred Fuchs Prize, the promotion award of the VG Wort for her Ph.D., the Young Scholars Award of the Research Network on EU Administrative Law, and the Heinz Maier Leibnitz Prize 2020 for her interdisciplinary research.

Oliver Mueller is a professor of philosophy with a focus on technology and on contemporary philosophy, University of Freiburg (Germany) and was a Senior Fellow at the 2018–2021 FRIAS Saltus Research Group 'Responsible AI – Emerging ethical, legal, philosophical, and social aspects of the interaction between humans and autonomous intelligent systems'. His areas of research are *inter alia* philosophy of technology (AI, bio-/neurotechnologies, human–machine interactions, digitalization), and philosophical anthropology, ethics. Since 2008, *Oliver Mueller* has been Principal Investigator in several interdisciplinary projects on current technologies. In 2015 he obtained the Heisenberg Grant (DFG) and a deputy professorship at the University of Koblenz-Landau, and between 2015 and 2017 he held a visiting professorship at the ETH, Zurich (Switzerland).

Richard Ngo is a student at the University of Cambridge. His dissertation focuses on researching the philosophy of machine learning, especially comparing AI development to the evolution of human intelligence. In 2017, he graduated with a first-class honours degree Bachelor in Computer Science and Philosophy from the University of Oxford (UK). He graduated with distinction, earning a Master of Philosophy from the University of Cambridge (UK). Prior to beginning his dissertation, *Richard Ngo* worked as a research engineer at DeepMind. There, he extensively investigated the foundations of AI safety research. His work experience also includes a summer internship at the Future of Humanity Institute.

Boris P. Paal is a professor of civil law and information law, data and media law, and a Director of the Institute for Information Law, Media Law, and Law of Digitalization at the University of Leipzig (Germany). Before this, *Boris Paal* was Director of the Institute for Media and Information Law, Dept. I (Civil Law), University of Freiburg (Germany). He researches and teaches, advises, and publishes in all areas of Private and Business Law with a special focus on data, competition, media, and information law as well as on compliance topics. As an author and editor (among others of Gersdorf/Paal on *Information and Media Law*, and Paal/Pauly on the *GDPR*), he is responsible for more than 180 publications in the aforementioned fields. He regularly works as an expert advisor for both state institutions and private entities.

Sangchul Park is an assistant professor at the Seoul National University School of Law (Korea). He completed his JSD at the University of Chicago and his undergraduate studies at Seoul National University. His main research area is the application of machine learning to legal studies. Prior to beginning his academic career, he spent more than 13 years in private practice specialising in technology, media, and telecommunications.

Matthias Paul was appointed as a professor for Digital Business Management at the Baden-Wuerttemberg Corporate State University (DHBW) in Loerrach (Germany) in 2019, after teaching for a couple of years at other business schools. Before his academic tenure, he served in senior management roles in the financial service industry, as CEO of Ned Davis Research in Boston (USA), an international investment research firm for the asset management industry, and in various managing director roles for financial market data vendors and information providers like IDC (today Factset) and Dow Jones, both in Germany and the USA. He started his business career as a Strategy Consultant at the Boston Consulting Group, working on digitalisation

projects across different industries, embracing the first internet revolution. His academic background is in AI, Cognitive Science, and analytical philosophy. He completed his Ph.D. at the University of Edinburgh at the Centre for Cognitive Science with a thesis on natural language understanding. His dissertation *Success in Referential Communication* was published in 1999 (reprint 2010) by Kluwer as part of the Philosophical Studies Series (Vol. 80).

Ralf Poscher is a Director at the Max Planck Institute for the Study of Crime, Security and Law in Freiburg and an Honorary Professor at the University of Freiburg (Germany). His research focuses on constitutional law, national security law, and legal theory.

Mathias Risse is Berthold Beitz Professor in Human Rights, Global Affairs, and Philosophy and Director of the Carr Center for Human Rights Policy at Harvard University (USA). His work primarily addresses questions of global justice ranging from human rights, inequality, taxation, trade and immigration to climate change, obligations to future generations, and the future of technology, especially the impact of AI on a range of normative issues. He has also worked on questions in ethics, decision theory, and nineteenth century German philosophy, especially *Nietzsche*. *Mathias Risse* is the author of *On Global Justice* (Princeton University Press) and *Global Political Philosophy* (Palgrave Macmillan), as well as *On Trade Justice: A Philosophical Plea for a New Global Deal* (Oxford University Press, with Gabriel Wollner) and *On Justice: Philosophy, History, Foundations* (Cambridge University Press).

Thorsten Schmidt is a professor of Mathematical Stochastics at the University of Freiburg (Germany). His research combines financial mathematics with the area of stochastic processes and statistics. In his career, he met interesting problems, both from statistics and financial mathematics, which inspire deeper mathematical understanding by their surprising complexity. In Freiburg, he and his research team are working on tackling these challenges with an improved mathematical model and targeting various applied areas where this can be useful. Besides finance, this includes medicine, robotics, and, in general, all areas where stochastic modelling is used.

Weixing Shen is a professor and Dean of the School of Law, Tsinghua University, Beijing (China), awarded with the 'National Outstanding Young Jurist' in 2014. He also serves as Executive Council Member of China Civil Law Society, Vice President of China Cyber and Information Law Society, etc. His research interests include civil law, computational law, and health law. He also hosts the National Key R&D Program 'Research on Intelligent Assistive Technology in the Context of Judicial Process Involving Concerned Civil and Commercial Cases' and the Major Project of National Social Science Foundation 'Research on the Legal-based Governance of Internet Economy'.

Jaan Tallinn is a founding engineer of Skype and Kazaa as well as a co-founder of the Cambridge Centre for the Study of Existential Risk (UK) and Future of Life Institute (Boston, USA). He is on the Board of Sponsors of the Bulletin of the Atomic Scientists, and has served on the High-Level Expert Group on AI at the European Commission, as well as on the Estonian President's Academic Advisory Board. *Jaan Tallinn* is also an active angel investor, a partner at Ambient Sound Investments, and a former Investor Director of the AI company DeepMind.

Johanna Thoma is an associate professor in the Department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science (UK). Before that, she completed her Ph.D. in Philosophy at the University of Toronto and was a visiting researcher at Stanford University and LMU Munich. *Johanna Thoma* has published widely in the areas of

practical rationality and decision theory, economic methodology, and ethics and public policy. She has a special interest in decision making under uncertainty and over time, by policymakers, individuals, and artificial agents.

Stefan Thomas is a professor at the Law Faculty of the Eberhard Karls University of Tübingen (Germany), where he holds the chair in Private Law, Commercial Law, Competition and Insurance Law since 2009. He is Director of the Tübingen Research Institute on the Determinants of Economic Activity (TRIDEA), and a Member of the International Advisory Board of the Institute for Global Law, Economics and Finance, Queen Mary University of London. His principal areas of research are German, European, and international antitrust- and competition law and regulation. His most recent work on the intersection between AI and antitrust is *Harmful Signals: Cartel Prohibition and Oligopoly Theory in the Age of Machine Learning* (2019), which was nominated for the antitrust writing awards 2020.

Antje von Ungern-Sternberg is a professor of Comparative Public Law and International Law at Trier University (Germany). She is a Director of the Institute for Legal Policy at the University of Trier and a Director of the Institute for Digital Law Trier. Her research focuses on comparative constitutional law, European and public international law, and the law of digitalisation.

Silja Voeneky is a professor of public international law, comparative law, and ethics of law at the University of Freiburg (Germany). She was a Fellow at Harvard Law School (2015–2016) and a Senior Fellow at the 2018–2021 FRIAS Saltus Research Group 'Responsible AI – Emerging ethical, legal, philosophical, and social aspects of the interaction between humans and autonomous intelligent systems'. Her areas of research include environmental law, laws of war, human rights, and the governance of disruptive research and technologies. She previously served as a Director of a Max Planck Research Group at the Max Planck Institute for Comparative Public and International Law in Heidelberg (2005–2010), and a member of the German Ethics Council appointed by the Federal Government. For many years, *Silja Voeneky* has been a legal advisor to the German Federal Foreign Office, and the German Federal Ministry of Environment. She is an author and (co-)editor of several articles and books in the fields of international law and the governance of emerging technologies, as for instance Voeneky/Neuman (eds), *Human Rights, Democracy, and Legitimacy in a World of Disorder* (Cambridge University Press 2018).

Christiane Wendehorst has been professor of civil law at the University of Vienna (Austria) since 2008. Amongst other functions, she is a founding member and Scientific Director of the European Law Institute (ELI), chair of the Academy Council of the Austrian Academy of Sciences (ÖAW) and co-head of the Department of Innovation and Digitalisation in Law. She is a member of the Bioethics Commission at the Austrian Federal Chancellery; a member of the Managing Board of the Austrian Jurists' Association (ÖJT); a member of the Academia Europea (AE), the International Academy for Comparative Law (IACL), and the American Law Institute (ALI). Currently, her research focuses on legal aspects of digitalisation and she has been working as an expert on topics such as digital content, the Internet of Things, AI, and data economy, for the European Commission, the European Parliament, the German Federal Government, the ELI and the ALI. *Christiane Wendehorst* is currently leading the transatlantic project 'Principles for a Data Economy' as well as various projects in the area of digitalisation, on topics such as safety- and liability-related aspects of software and biometric techniques. She is also involved in a number of projects on algorithmic fairness. Prior to moving to Vienna, she was a professor in Göttingen (1999–2008) and Greifswald (1998–1999) and was Managing Director of the Sino-German Institute of Legal Studies (2000–2008).

Jonathan Zittrain is the George Bemis Professor of International Law at Harvard Law School (USA). He is also a professor at the Harvard Kennedy School of Government, a professor of computer science at the Harvard School of Engineering and Applied Sciences, Director of the Harvard Law School Library, and co-founder and Director of Harvard's Berkman Klein Center for Internet & Society. His research interests include the ethics and governance of AI; battles for control of digital property; the regulation of cryptography; new privacy frameworks for loyalty to users of online services; the roles of intermediaries within Internet architecture; and the useful and unobtrusive deployment of technology in education. *Jonathan Zittrain* established and co-leads the Institute for Rebooting Social Media, a three-year 'pop-up' research initiative that is bringing together participants from across industry, governments, and academia in a focused, time-bound collaboration. He also championed the development of the Caselaw Access Project, which has expanded free public access to US case law. His book, *The Future of the Internet – And How to Stop It*, predicted the end of general purpose client computing and the corresponding rise of new gatekeepers.

Acknowledgements

We want to start by thanking the Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, for the immense support of our Saltus Responsible AI Research Group 2018–2021.* Without the FRIAS, its directors, especially Bernd Kortmann, and the team, this Handbook would not have been possible.

We would also like to thank the researchers at our Responsible AI Research Group and at the chair of Silja Voeneky, Alisa Pojtinger, Jonatan Klaedtke, and Daniel Feuerstack, who helped us edit this book, for their important work. Furthermore, we are very grateful to the research assistants Alena Huenermund (Responsible AI Research Group), Hannah Weber, and Heidi He (University of Cambridge) for their valuable support.

We would also like to acknowledge the Baden-Württemberg Stiftung for their financial support in the context of our research project AI Trust. Finally, we would like to express our thanks to the Klaus Tschira Foundation for supporting the work of editor Philipp Kellmeyer in the project "Neuroethics and AI Ethics Lab" (grant no. 00.001.2019)."

* UP 14/1 (2), DFG-ProjektID: 422010107.

Introduction

Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard

In the past decade, Artificial Intelligence (AI) as a general-purpose tool has become a disruptive force globally. By leveraging the power of artificial neural networks, deep learning frameworks can now translate text from hundreds of languages, enable real-time navigation for everyone, recognise pathological medical images, as well as enable many other applications across all sectors in society. However, the enormous potential for innovation and technological advances and the chances that AI systems provide come with hazards and risks that are not yet fully explored, let alone fully understood. One can stress the opportunities of AI systems to improve healthcare, especially in times of a pandemic, provide automated mobility, support the protection of the environment, protect our security, and otherwise support human welfare. Nevertheless, we must not neglect that AI systems can pose risks to individuals and societies; for example by disseminating biases, by undermining political deliberation, or by the development of autonomous weapons. This means that there is an urgent need for responsible governance of AI systems. This Handbook shall be a basis to spell out in more detail what could become relevant features of Responsible AI and how we can achieve and implement them at the regional, national, and international level. Hence, the aim of this Handbook is to address some of the most pressing philosophical, ethical, legal, and societal challenges posed by AI.

But mapping the uncertainties, benefits, and risks of AI systems in general and with regard to different sectors and assessing relevant ethical and legal rules requires a wide range of expertise from areas such as computer science, robotics, mathematical modelling, as well as trans- and interdisciplinary normative analyses from law, philosophy, and ethics by authors from different continents. Therefore, the authors of this Handbook explore the technical and conceptual foundations as well as normative aspects of Responsible AI from many different angles.

OUTLINE

The Handbook consists of eight parts and begins with the foundations of Responsible AI (first part), the current and future approaches to AI governance (second part) not limited to European and US approaches. Authors further analyse liability schemes (third part) and shed light on the core problem of fairness and non-discrimination in AI systems (fourth part) before approaches in responsible data governance are examined in more detail (fifth part). The authors of the sixth and seventh parts discuss justified governance approaches of specific sectors of AI systems: these systems can be part of such important fields as corporate governance, including financial

services, and the healthcare sector, including neurotechnology. Authors of the eighth part tackle especially problematic and challenging issues such as the use of AI for security applications and in armed conflict.

PART I: FOUNDATIONS

The first chapter, written by *Wolfram Burgard*, describes features of AI systems, introducing *inter alia* the notions of machine learning and deep learning as well as the use of AI systems as part of robotics. In doing so, *Burgard* provides an overview of the technological state of the art from the perspective of robotics and computer science.

Jaan Tallinn and *Richard Ngo* propose an answer to the question of what Responsible AI means by offering a framework for the deployment of AI which focuses on two concepts: delegation and supervision. This framework aims towards building 'delegate AIs' which lack goals of their own but can perform any task delegated to them. However, AIs trained with hardcoded reward functions, or even human feedback, often learn to game their reward signal instead of accomplishing their intended tasks. Thus, *Tallinn* and *Ngo* argue that it will be important to develop more advanced techniques for continuous, high-quality supervision; for example, by evaluating the reasons which AI-systems give for their choices of actions. These supervision techniques might be made scalable by training AI-systems to generate reward signals for more advanced AI-systems.

After this, the philosopher *Catrin Misselhorn* provides an overview of the core debates in artificial morality and machine ethics. She argues that artificial moral agents are AI systems which are able to recognise morally relevant factors and take them into account in their decisions and actions. *Misselhorn* shows that artificial morality is not just a matter of science fiction scenarios but rather an issue that has to be considered today. She provides a basis for what Responsible AI means by laying down conceptual foundations of artificial morality and discussing related ethical issues.

The philosopher *Johanna Thoma* focuses, as part of the fourth chapter, on the 'moral proxy problem,' which arises when an autonomous artificial agent makes a decision as a proxy for a human agent without it being clear for whom specifically it does so. *Thoma* identifies two categories of agents an artificial agent can be a proxy for: *low-level* agents (individual users or the kinds of human agents that are typically replaced by artificial agents) and *high-level* agents (designers, distributors, or regulators). She shows that we do not get the same recommendations under different agential frames: whilst the former suggests the agents be programmed without risk neutrality, which is common in choices made by humans, the latter suggests the contrary, since the choices are considered part of an aggregate of many similar choices.

In the final chapter of the first part, the philosopher *Christoph Durt* develops a novel view on AI and its relation to humans. *Durt* contends that AI is neither merely a tool, nor an artificial subject, nor necessarily a simulation of human intelligence, but rather AI is defined by its *interrelational character*. According to this author, misconceptions of AI have led to far-reaching misunderstandings of the opportunities and risks of AI. Hence, a more comprehensive concept of AI is needed to better understand the possibilities of Responsible AI. He argues that the setup of the *Turing Test* is already deceptive, as this test avoids difficult philosophical questions by passing on the burden to an evaluator, and delineates a more comprehensive picture according to which AI integrates into the human lifeworld through its interrelations with humans and data.

PART II: APPROACHES TO AI GOVERNANCE

The second part of the Handbook, which focuses on current and future approaches to AI governance, begins with a chapter by the philosopher *Mathias Risse*. He reflects on the medium- and long-term prospects and challenges democracy faces from AI. *Risse* argues that both technologists and citizens need to engage with ethics and political thoughts in order to build and maintain a democracy-enhancing AI infrastructure, and stresses the need for Responsible AI as he points out AI's potential to greatly strengthen democracy, but only with the right efforts. His answer starts with a critical examination of the relation between democracy and technology with a historical perspective before outlining a 'techno skepticism' prevalent in several grand narratives of AI. *Risse* explores the possibilities and challenges that AI may lead to and argues that technology critically bears on what forms of human life get realised or imagined. According to the author, AI changes the materiality of democracy by altering how collective decision making unfolds and what its human participants are like.

Next, *Thomas Burri*, an international law scholar, examines how general ethical norms on AI diffuse into domestic law without engaging international law. *Burri* discusses various frameworks for 'ethical AI' and shows how they influenced the European Union Commission's proposal for the draft 2021 AI Act. Hence, this chapter reveals the origins of the EU proposal and explains the substance of the future EU AI regulation.

From an interdisciplinary angle, the mathematician *Thorsten Schmidt* and the ethics and international law scholar *Silja Voeneky* propose a new adaptive regulation scheme for AI-driven high-risk products and services as part of a future Responsible AI governance regime. They argue that current regulatory approaches with regard to these products and services, including the draft 2021 EU AI Act, have to be supplemented by a new regulatory approach. At its core, the adaptive AI regulation proposed by the authors requires that private actors, like companies developing and selling high-risk AI-driven products and services, pay a proportionate amount of money as a financial guarantee into a fund before the AI-based high-risk product or service enters the market. *Schmidt* and *Voeneky* spell out in more detail what amount of regulatory capital can be seen as proportionate and what kind of accompanying rules are necessary to implement this adaptive AI regulation.

In chapter nine, the law scholars *Weixing Shen* and *Yun Liu* focus on China's AI regulation. They show that there is no unified AI law today in China but argue that many provisions from Chinese data protection law are in part applicable to AI systems. The authors particularly analyse the rights and obligations from the Chinese Data Security Law, the Chinese Civil Code, the E-Commerce Law, and the Personal Information Protection Law; they finally introduce the Draft Regulation on Internet Information Service Based on Algorithm Recommendation Technology and explain the relevance of these regulations with regard to algorithm governance.

Current and future approaches to AI governance have to be looked at from a philosophical perspective, too, and *Thomas Metzinger* lists the main problem domains related to AI systems from a philosophical angle. For each problem domain, he proposes several measures which should be taken into account. He starts by arguing that there should be worldwide safety standards concerning the research and development of AI. Additionally, a possible AI arms race must be prevented as early as possible. Thirdly, he stresses that any creation of artificial consciousness should be avoided, as it is highly problematic from an ethical point of view. Besides, he argues that synthetic phenomenology could lead to non-biological forms of suffering and might lead to a vast increase of suffering in the universe, as AI can be copied rapidly, and that in the field of AI systems there is the risk of unknown risks.

In the final chapter of this part, the law and technology scholar *Jonathan Zittrain* begins with the argument that there are benefits to utilising scientific solutions discovered through trial and error rather than rigorous proof (though aspirin was discovered in the late nineteenth century, it was not until the late twentieth century that scientists were able to explain how it worked), but argues that doing so accrues 'intellectual debt'. This intellectual debt is compounding quickly in the realm of AI, especially in the subfield of machine learning. Society's movement from basic science towards applied technology that bypasses rigorous investigative research inches us closer to a world in which we are reliant on an oracle AI, one in which we trust regardless of our ability to audit its trustworthiness. *Zittrain* concludes that we must create an intellectual debt 'balance sheet' by allowing academics to scrutinise the systems.

PART III: RESPONSIBLE AI LIABILITY SCHEMES

The next part of the Handbook focuses on Responsible AI liability schemes from a legal perspective. First of all, *Christiane Wendehorst* analyses the different potential risks posed by AI as part of two main categories, safety risks and fundamental rights risks, and considers why AI challenges existing liability regimes. She highlights the fact that liability for fundamental rights risks is largely unchartered while being AI-specific. *Wendehorst* argues that a number of changes have to be made for the emerging AI safety regime to be used as a 'backbone' for the future AI liability regime if this is going to help address liability for fundamental rights risks. As a result, she suggests that further negotiations about the AI Act proposed by the European Commission should be closely aligned with the preparatory work on a future AI liability regime.

Secondly, the legal scholar *Jan von Hein* analyses and evaluates the European Parliament's proposal on a civil liability regime for AI against the background of the already existing European regulatory framework on private international law, in particular the Rome I and II Regulations. He argues that the draft regulation proposed by the European Parliament is noteworthy from a private international law perspective because it introduces new conflicts rules for AI. In this regard, the proposed regulation distinguishes between a rule delineating the spatial scope of its autonomous rules on strict liability for high-risk AI systems, on the one hand, and a rule on the law applicable to fault-based liability for low-risk systems, on the other hand. *Von Hein* concludes that, compared with the current Rome II Regulation, the draft regulation would be a regrettable step backwards.

PART IV: FAIRNESS AND NON-DISCRIMINATION

The fourth part of the Handbook, which links fairness and non-discrimination in AI systems to the concept of Responsible AI, begins with a chapter by the philosopher *Wilfried Hinsch*. He focuses on statistical discrimination by means of computational profiling. He argues that because AI systems do not rely on human stereotypes or rather limited data, computational profiling may be a better safeguard of fairness than humans. He starts by defining statistical profiling as an estimate of what individuals will do by considering the group of people they can be assigned to, and explores which criteria of fairness and justice are appropriate for the assessment of computational profiling. *Hinsch* argues that discrimination constitutes a rule-guided social practice that imposes unreasonable burdens on specific people. He spells out that even statistically correct profiles can be unacceptable considering reasons of procedural fairness or substantive justice. Because of this, he suggests a fairness index for profiles to determine procedural fairness.

In Chapter 15, the legal scholar *Antje von Ungern-Sternberg* focuses on the legality of discriminatory AI, which is increasingly used to assess people (profiling). As many studies show that the use of AI can lead to discriminatory outcomes, she aims to answer the question of whether the law as it stands prohibits objectionable forms of differential treatment and detrimental impact. She takes up the claim that we need a 'right to reasonable inferences' with respect to discriminatory AI and argues that such a right already exists in antidiscrimination law. Also, *von Ungern-Sternberg* shows that the need to justify differential treatment and detrimental impact implies that profiling methods correspond to certain standards, and that these methodological standards have yet to be developed.

PART V: RESPONSIBLE DATA GOVERNANCE

The fifth part of the Handbook analyses problems of responsible data governance. The legal scholar *Ralf Poscher* sets out to show, in Chapter 16, how AI challenges the traditional understanding of the right to data protection and presents an outline of an alternative conception that better deals with emerging AI technologies. He argues that we have to step back from the idea that each and every instance of personal data processing concerns a fundamental right. For this, *Poscher* explains how the traditional conceptualisation of data protection as an independent fundamental right on its own collides with AI's technological development, given that AI systems do not provide the kind of transparency required by the traditional approach. And secondly, he proposes an alternative model, a no-right thesis, which shifts the focus from data protection as an independent right to other existing fundamental rights, such as liberty and equality.

The legal scholar *Boris Paal* also identifies a conflict between two objectives pursued by data protection law, the comprehensive protection of privacy and personal rights and the facilitation of an effective and competitive data economy. Focusing on the European Union's General Data Protection Regulation (GDPR), the author recognises its failure to address the implications of AI, the development of which depends on the access to large amounts of data. In general, he argues, that the main principles of the GDPR seem to be in direct conflict with the functioning and underlying mechanisms of AI applications, which evidently were not considered sufficiently whilst the regulation was being drafted. Hence, *Paal* argues that establishing a separate legal basis governing the permissibility of processing operations using AI-based applications should be considered.

In the last chapter of the fifth part, the legal scholars *Sangchul Park*, *Yong Lim*, and *Haksoo Ko*, analyse how South Korea has been dealing with the COVID-19 pandemic and its legal consequences. Instead of enforcing strict lockdowns, South Korea imposed a robust, AI-based contact tracing scheme. The chapter provides an overview of the legal framework and the technology which allowed the employment of this technology-based contact tracing scheme. The authors argue that South Korea has a rather stringent data-protection regime, which proved to be the biggest hurdle in implementing the contact tracing scheme. However, the state introduced a separate legal framework for extensive contact tracing in 2015, which was reactivated and provided government agencies with extensive authority to process personal data for epidemiological purposes.

PART VI: RESPONSIBLE CORPORATE GOVERNANCE OF AI SYSTEMS

The sixth part looks at responsible corporate governance of AI systems and it starts by exploring the changes that AI brings about in corporate law and corporate governance. The legal scholar

Jan Lieder argues that whilst there is the potential to enhance the current system, there are also risks of destabilisation. Although algorithms are already being used in the board room, lawmakers should not consider legally recognizing e-persons as directors and managers. Rather, scholars should evaluate the effects of AI on the corporate duties of boards and their liabilities. Finally, *Lieder* suggests the need for transparency in a company's practices regarding AI for awareness-raising and the enhancement of overall algorithm governance, as well as the need for boards to report on their overall AI strategy and ethical guidelines relating to the responsibilities, competencies, and protective measures they established.

In Chapter 20, it is shown how enforcement paradigms that hinge on descriptions of the inner sphere and conduct of human beings may collapse when applied to the effects precipitated by independent AI-based computer agents. It aims to serve as a conceptual sketch for the intricacies involved in autonomous algorithmic collusion, including the notion of concerted practices for cases that would otherwise elude the cartel prohibition. *Stefan Thomas*, a legal scholar, starts by assessing how algorithms can influence competition in markets before dealing with the traditional criteria of distinction between explicit and tacit collusion. This might reveal a potential gap in the existing legal framework regarding algorithmic collusion. Finally, *Thomas* analyses whether the existing cartel prohibition can be construed in a manner that captures the phenomenon appropriately.

Matthias Paul explores in the next chapter the topic of AI systems in the financial sector. After outlining different areas of AI application and different regulatory regimes relevant to robo-finance, he analyses the risks emerging from AI applications in the financial industry. The author argues that AI systems applied in this sector usually do not create new risks. Instead, existing risks can actually be mitigated through AI applications. *Paul* analyses personal responsibility frameworks that have been suggested by scholars in the field of robo-finance, and shows why they are not a sufficient approach for regulation. He concludes by discussing the draft 2021 EU AI Act as a suitable regulatory approach based on the risks linked to specific AI systems and AI-based practices.

PART VII: RESPONSIBLE AI IN HEALTHCARE AND NEUROTECHNOLOGY

As another important AI-driven sector, the seventh part of the Handbook focuses on Responsible AI in healthcare and neurotechnology governance. The legal scholars *Fruzsina Molnár-Gábor* and *Johanne Giesecke* begin by setting out the key elements of medical AI. They consider the aspects of how the application of AI-based systems in medical contexts may be guided according to international standards. The authors argue that among the frameworks that exist, the World Medical Association's codes appear particularly promising as a guide for standardisation processes. *Molnár-Gábor* and *Giesecke* sketch out the potential applications of AI and its effects on the doctor–patient relationship in terms of information, consent, diagnosis, treatment, aftercare, and education.

In Chapter 23, the legal scholar *Christoph Krönke* focuses on the legal challenges healthcare AI Alter Egos face, especially in the EU, as these AI Alter Egos have two main functions, collecting a substantive database and proposing diagnoses. The author spells out the relevance of European data protection laws and analyses the European Medical Devices Regulation (MDR) with regard to the responsible governance of these entities. *Krönke* argues that AI Alter Egos are regulated by an appropriate legal framework in the EU today, but it nevertheless has to be open for developments in order to remain appropriate.

In the following chapter, neurologist and neuroethics scholar *Philipp Kellmeyer* sets out a human-rights based approach for governing AI-based neurotechnologies. *Kellmeyer* outlines the

current scholarly discussion and policy initiatives about neurorights and discusses how to protect mental privacy and mental integrity. He argues that mental privacy and integrity are important anthropological goods that need to be protected from unjustified interferences. He argues that while existing human rights provide a sufficient legal basis, an approach is required that makes these rights actionable and justiciable to protect mental privacy and mental integrity.

In the final chapter of this part, the philosophers *Boris Essmann* and *Oliver Mueller* address AI-supported neurotechnology, especially Brain–Computer Interfaces (BCIs) that may in the future supplement and restore functioning in agency-limited individuals or even augment or enhance capacities for natural agency. The authors propose a normative framework for evaluating neurotechnological and AI-assisted agency based on 'cyberbilities' that can be part of a Responsible AI framework. 'Cyberbilities' are capabilities that emerge from human–machine interactions in which agency is distributed across human and artificial elements. *Essmann* and *Mueller* suggest a list of 'cyberbilities' that is meant to support the well-being of individuals.

PART VIII: RESPONSIBLE AI FOR SECURITY APPLICATIONS AND IN ARMED CONFLICT

The eighth and final part of this Handbook discusses the highly controversial use of Responsible AI for security applications and in armed conflict. The legal scholar *Ebrahim Afsah* outlines different implications of AI for the area of national security. He argues that while AI overlaps with many challenges to the national security arising from cyberspace, it also creates new risks, including the development of autonomous weapons, the enhancement of existing military capabilities, and threats to foreign relations and economic stability. Most of these risks, however, *Afsah* argues, can be subsumed under existing normative frameworks.

In the next chapter, the political philosopher *Alex Leveringhaus* spells out ethical concerns regarding autonomous weapons systems (AWS) by asking whether lethal autonomous weapons are morally repugnant and whether this entails that they should be prohibited by international law. *Leveringhaus* surveys three prominent ethical arguments against AWS: firstly, autonomous weapons systems create 'responsibility gaps'; secondly, that their use is incompatible with human dignity; and, thirdly, that autonomous weapons systems replace human agency with artificial agency. He argues that some of these arguments fail to show that autonomous weapons systems are morally different from more established weapons. However, the author concludes that autonomous weapons systems are problematic due to their lack of predictability.

In the final chapter of this Handbook, the legal scholar *Dustin Lewis* discusses the use of Responsible AI during armed conflict. The scope of this chapter is not limited to lethal autonomous weapons but also encompasses other AI-related tools and techniques related to warfighting, detention, and humanitarian services. For this, he explores the requirements of international law and outlines some preconditions necessary to respect international law. According to *Lewis*, current international law essentially presupposes humans as legal agents. From that premise, the author argues that any employment of AI-related tools or techniques in an armed conflict needs to be susceptible to being administered, discerned, attributed, understood, and assessed by human agents.

After this outline about the core issues discussed with regard to the concept of Responsible AI, we want to note that many chapters of the Handbook have strong links to an international and interdisciplinary virtual research conference on "*Global Perspectives on Responsible AI*". We convened this conference in June 2020 based at the *Freiburg Institute for Advanced Studies* and edited this Handbook in order to shed more light on the transformations that are based on the rise

of AI systems, their impact on our societies, and the challenges for the responsible governance and regulation of AI. Although the chapters of this Handbook shall not – and cannot – answer all questions with regard to AI governance and more problems have to be discussed and solved in the forthcoming years, we as editors agree that Responsible AI governance shall be conducive to scientific and technological progress, to our stability and flourishing as individuals and as humanity. We hope that the perspectives, analyses, and proposals for a concept of Responsible AI in this Handbook will provide a basis for fostering deliberations to develop and spell out proportional approaches to AI governance and enable scholars, scientists, and other actors to discuss normative frameworks for AI that allow societies, states, and the international community to unlock the potential for responsible innovation in this important field.

PART I

Foundations of Responsible AI

1

Artificial Intelligence

Key Technologies and Opportunities

Wolfram Burgard

I. INTRODUCTION

Artificial Intelligence (AI) is a discipline that is concerned with the generation of software systems that provide functions, the execution of which requires what is typically referred to by the word intelligence. Thereby, the corresponding tasks can be performed by pure software agents as well as by physical systems, such as robots or self-driving cars.

As the term 'intelligence' is already very difficult to define, the definition of AI is, of course, correspondingly difficult and numerous definitions can be found in the literature.[1] Among them are several approaches that are based on human behavior or thinking. For example, the *Turing* test[2] introduced by *Alan Turing* in 1950, in which the actions generated by the system or robot should not be distinguishable from those generated by humans, has to be mentioned in this context. Such a *Turing* test for systems interacting with humans would then mean, for example, that a human could no longer determine whether a conversation partner on the telephone is a human or software.

However, most current AI systems aim to generate agents that think or act rationally. To realize systems that think rationally, often logic-based representations and reasoning systems are used. The basic assumption here is that rational thinking entails rational action if the reasoning mechanisms used are correct. Another group of definitional approaches deals with the direct generation of rational actions. In such systems, the underlying representations often are not human-readable or easily understood by humans. They often use a goal function that describes the usefulness of states. The task of the system is then to maximize this objective function, that is, to determine the state that has the maximum usefulness or that, in case of uncertainties, maximizes the future expected reward. If, for example, one chooses the cleanliness of the work surface minus the costs for the executed actions as the objective function for a cleaning robot, then in the ideal case this leads to the robot selecting the optimal actions in order to keep the work surface as clean as possible. This already shows the strength of the approach to generate rational behavior compared to the approach to generate human behavior. A robot striving for rational behavior can simply become more effective than one that merely imitates human behavior, because humans, unfortunately, do not show the optimal behavior in all cases. The disadvantage lies in the fact that the interpretation of the representations or structures learned by

[1] NJ Nilsson, *Artificial Intelligence: A New Synthesis* (1998); S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (4th ed. 2016).

[2] A Turing, 'Computing Machinery and Intelligence' (1950) 59 *Mind* 433.

the system typically is not easy, which makes verification difficult. Especially in the case of safety-relevant systems, it is often necessary to provide evidence of the safety of, for example, the control software. However, this can be very difficult and generally even impossible to do analytically, so one has to rely on statistics. In the case of self-driving cars, for example, one has to resort to extensive field tests in order to be able to prove the required safety of the systems.

Historically, the term AI dates back to 1956, when at a summer workshop called the Dartmouth Summer Research Project on Artificial Intelligence,[3] renowned scientists met in the state of New Hampshire, USA, to discuss AI. The basic idea was that any aspect of learning or other properties of intelligence can be described so precisely that machines can be used to simulate them. In addition, the participants wanted to discuss how to get computers to use language and abstract concepts, or simply improve their own behavior. This meeting is still considered today to have been extremely successful and has led to a large number of activities in the field of AI. For example, in the 1980s, there was a remarkable upswing in AI in which questions of knowledge representation and knowledge processing played an important role. In this context, for example, expert systems became popular.[4] Such systems used a large corpus of knowledge, represented for example in terms of facts and rules, to draw conclusions and provide solutions to problems. Although there were initially quite promising successes with expert systems, these successes then waned quite a bit, leading to a so-called demystification of AI and ushering in the AI winter.[5] It was not until the 1990s when mathematical and probabilistic methods increasingly took hold and a new upswing could be recorded. A prominent representative of this group of methods is Bayesian networks.[6] The systems resulting from this technique were significantly more robust than those based on symbolic techniques. This period also started the advent of machine learning techniques based on probabilistic and mathematical concepts. For example, support vector machines[7] revolutionized machine learning. Until a few years ago, they were considered one of the best performing approaches to classification problems. This radiated to other areas, such as pattern recognition and image processing. Face recognition and also speech recognition algorithms found their way into products we use in our daily lives, such as cameras or even cell phones. Cameras can automatically recognize faces and cell phones can be controlled by speech. These methods have been applied in automobiles, for example when components can be controlled by speech. However, there are also fundamental results from the early days of AI that have a substantial influence on today's products. These include, for example, the ability of navigation systems to plan the shortest possible routes[8] and navigate us effectively to our destination based on given maps. Incidentally, the same approaches play a significant role in computer games, especially when it comes to simulating intelligent systems that can effectively navigate the virtual environment. At the same time, there was also a paradigm shift in robotics. The probabilistic methods had a significant impact, especially on the navigation of mobile robots, and today, thanks to this development, it is well understood how to build mobile systems that move autonomously in their environment. This currently has an

[3] J McCarthy and others, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955' (2006) 27(4) *AI Magazine* 12.

[4] F Hayes-Roth, DA Waterman, and DB Lenat, *Building Expert Systems* (1983).

[5] E Fast and E Horvitz, 'Long-Term Trends in the Public Perception of Artificial Intelligence' (2017) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI).

[6] J Pearl, *Causality: Models, Reasoning and Inference* (2009) (hereafter Pearl, *Causality*).

[7] VN Vapnik, *Statistical Learning Theory* (1998) (hereafter Vapnik, *Statistical Learning Theory*).

[8] PE Hart, NJ Nilsson, and B Raphael, 'A Formal Basis for the Heuristic Determination of Minimum Cost Paths' (1968) 4(2) *IEEE Transactions on Systems Science and Cybernetics* 100 (hereafter Hart and others, 'A Formal Basis for the Heuristic Determination of Minimum Cost Paths').

important influence on various areas, such as self-driving cars or transport systems in logistics, where extensive innovations can be expected in the coming years.

For a few years now, the areas of machine learning and robotics have been considered particularly promising, based especially on the key fields of big data, deep learning, and autonomous navigation and manipulation.

II. MACHINE LEARNING

Machine learning typically involves developing algorithms to improve the performance of procedures based on data or examples and without explicit programming.[9] One of the predominant applications of machine learning is that of classification. Here the system is presented with a set of examples and their corresponding classes. The system must now learn a function that maps the properties or attributes of the examples to the classes with the goal of minimizing the classification error. Of course, one could simply memorize all the examples, which would automatically minimize the classification error, but such a procedure would require a lot of space and, moreover, would not generalize to examples not seen before. In principle, such an approach can only guess. The goal of machine learning is rather to learn a compact function that performs well on the given data and also generalizes well to unseen examples. In the context of classification, examples include decision trees, random forests, a generalization thereof, support vector machines, or boosting. These approaches are considered supervised learning because the learner is always given examples including their classes.

Another popular supervised learning problem is regression. Here, the system is given a set of points of a function with the task of determining a function that approximates the given points as well as possible. Again, one is interested in functions that are as compact as possible and minimize the approximation error. In addition, there is also unsupervised learning, where one searches for a function that explains the given data as well as possible. A typical unsupervised learning problem is clustering, where one seeks centers for a set of points in the plane such that the sum of the squared distances of all points from their nearest center is minimized.

Supervised learning problems occur very frequently in practice. For example, consider the face classification problem. Here, for a face found in an image, the problem is to assign the name of the person. Such data is available in large masses to companies that provide social networks, such as Facebook. Users can not only mark faces on Facebook but also assign the names of their friends to these marked faces. In this way, a huge data set of images is created in which faces are marked and labelled. With this, supervised learning can now be used to (a) identify faces in images and (b) assign the identified faces to people. Because the classifiers generalize well, they can subsequently be applied to faces that have not been seen before, and nowadays they produce surprisingly good results.

In fact, the acquisition of large corpora of annotated data is one of the main problems in the context of big data and deep learning. Major internet companies are making large-scale efforts to obtain massive corpora of annotated data. So-called CAPTCHAs (Completely Automated Public *Turing* tests to tell Computers and Humans Apart) represent an example of this.[10] Almost everyone who has tried to create a user account on the Internet has encountered such CAPTCHAs. Typically, service providers want to ensure that user accounts are not registered *en*

[9] TM Mitchell, *Machine Learning* (1997).

[10] L Von Ahn and others, 'CAPTCHA: Using Hard AI Problems for Security' (2003) Proceedings of the 22nd International Conference on Theory and Applications of Cryptographic Techniques, EUROCRYPT'03, 294.

masse by computer programs. Therefore, the applicants are provided with images of distorted text that can hardly be recognized by scanners and optical character recognition. Because the images are now difficult to recognize by programs, they are ideal for distinguishing humans from computer programs or bots. Once humans have annotated the images, learning techniques can again be used to solve these hard problems and further improve optical character recognition. At the same time, this ensures that computer programs are always presented with the hardest problems that even the best methods cannot yet solve.

1. *Key Technology Big Data*

In 2018, the total amount of storage globally available was estimated to be about 20 zettabytes (1 zettabyte = 10^{21} byte = 10^{9} terabytes).[11] Other sources estimate internet data transfer at approximately 26 terabytes per second.[12] Of course, predictions are always subject to large uncertainties. Estimates from the International Data Corporation assume that the total volume will grow to 160 zettabytes by 2025, an estimated tenfold increase. Other sources predict an annual doubling. The number of pages of the World Wide Web indexed by search engines is enormous. Google announced almost ten years ago that they have indexed 10^{12} different URLs (uniform resource locators, reference to a resource on the World Wide Web).[13] Even though these figures are partly based on estimates and should therefore be treated with caution, especially with regard to predictions for the future, they make it clear that huge amounts of data are available on the World Wide Web. This creates an enormous potential of data that is available not only to people but also to service providers such as Apple, Facebook, Amazon, Google, and many others, in order to offer services that are helpful to people in other contexts using appropriate AI methods. One of the main problems here, however, is the provision of data. Data is not helpful in all cases. As a rule, it only becomes so when people annotate it and assign a meaning to it. By using learning techniques, images that have not been seen before can be annotated. The techniques for doing so will be presented in the following sections. We will also discuss which methods can be used to generate this annotated data.

2. *Key Technology Deep Learning*

Deep learning[14] is a technique that emerged a few years ago and that can learn from massive amounts of data to provide effective solutions to a variety of machine learning problems. One of the most popular approaches is the so-called deep neural networks. They are based on the neural networks whose introduction dates back to *Warren McCulloch* and *Walter Pitts* in 1943.[15] At that time, they tried to reproduce the functioning of neurons of the brain by using electronic circuits, which led to the artificial neural networks. The basic idea was to build a network consisting of interconnected layers of nodes. Here, the bottom layer is considered the input layer, and the top

[11] D Reinsel, J Gantz, and J Rydning, 'Data Age 2025: The Evolution of Data to Life-Critical' (IDC White Paper, 2017) www.import.io/wp-content/uploads/2017/04/Seagate-WP-DataAge2025-March-2017.pdf.

[12] Ibid.

[13] J Alpert and N Hajaj, 'We knew the web was big. . .' (Google Blog, 2008) https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[14] I Arel, DC Rose, and TP Karnowski, 'Research Frontier: Deep Machine Learning – A New Frontier in Artificial Intelligence Research' (2010) 5(4) *IEEE Computational Intelligence Magazine* 1; Y LeCun, Y Bengio, and G Hinton, 'Deep Learning' (2015) 521 *Nature* 436.

[15] WS McCulloch and WH Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity' (1943) 5 *Bulletin of Mathematical Biophysics* 115.

layer is considered the output layer. Each node now executes a simple computational rule, such as a simple threshold decision. The outputs of each node in a layer are then passed to the nodes in the next layer using weighted sums. These networks were already extremely successful and produced impressive results, for example, in the field of optical character recognition. However, even then there were already pioneering successes from today's point of view, for example in the No Hands Across America project,[16] in which a minivan navigated to a large extent autonomously and controlled by a neural network from the east coast to the west coast of the United States. Until the mid-80s of the last century, artificial neural networks played a significant role in machine learning, until they were eventually replaced by probabilistic methods and, for example, Bayesian networks,[17] support vector machines,[18] or Gaussian processes.[19] These techniques have dominated machine learning for more than a decade and have also led to numerous applications, for example in image processing, speech recognition, or even human–machine interaction. However, they have recently been superseded by the deep neural networks, which are characterized by having a massive number of layers that can be effectively trained on modern hardware, such as graphics cards. These deep networks learn representations of the data at different levels of abstraction at each layer. Particularly in conjunction with large data sets (big data), these networks can use efficient algorithms such as backpropagation to optimize the parameters in a single layer based on the previous layer to identify structures in data. Deep neural networks have led to tremendous successes, for example in image, video, or speech processing. But they have also been used with great success in other tasks, such as in the context of object recognition or deep data interpretation. The deep neural networks could impressively demonstrate their ability in their application within AlphaGo, a computer program that defeated *Lee Sidol*, one of the best Go players in the world.[20] This is noteworthy because until a few years ago it was considered unlikely that Go programs would be able to play at such a level in the foreseeable future.

III. ROBOTICS

Robotics is a scientific discipline that deals with the design of physical agents (robotic systems) that effectively perform tasks in the real world. They can thus be regarded as physical AI systems. Application fields of robotics are manifold. In addition to classical topics such as motion planning for robot manipulators, other areas of robotics have gained increasing interest in the recent past, for example, position estimation, simultaneous localization and mapping, and navigation. The latter is particularly relevant for transportation tasks. If we now combine manipulators with navigating platforms, we obtain mobile manipulation systems that can play a substantial role in the future and offer various services to their users. For example, production processes can become more effective and also can be reconfigured flexibly with these robots. To build such systems, various key competencies are required, some of which are already available or are at a quality level sufficient for a production environment, which has significantly increased the attractiveness of this technology in recent years.

[16] C Thorpe and others, 'Toward Autonomous Driving: The CMU Navlab. I. Perception' (1991) 6(4) *IEEE Expert* 31.
[17] Pearl, *Causality* (n 6).
[18] Vapnik, *Statistical Learning Theory* (n 7).
[19] CE Rasmussen and CKI Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (2005).
[20] D Silver and others, 'Mastering the Game of Go with Deep Neural Networks and Tree Search' (2016) 529 *Nature* 484.

1. Key Technology Navigation

Mobile robots must be able to navigate their environments effectively in order to perform various tasks effectively. Consider, for example, a robotic vacuum cleaner or a robotic lawnmower. Most of today's systems do their work by essentially navigating randomly. As a result, as time progresses, the probability increases that the robot will have approached every point in its vicinity once so that the task is never guaranteed but very likely to be completed if one waits for a sufficiently long time. Obviously, such an approach is not optimal in the context of transport robots that are supposed to move an object from the pickup position to the destination as quickly as possible. Several components are needed to execute such a task as effectively as possible. First, the robot must have a path planning component that allows it to get from its current position to the destination point in the shortest possible path. Methods for this come from AI and are based, for example, on the well-known A* algorithm for the effective computation of shortest paths.[21] For path planning, robotic systems typically use maps, either directly in the form of roadmaps or by subdividing the environment of the robot into free and occupied space in order to derive roadmaps from this representation. However, a robot can only assume under very strong restrictions that the once planned path is actually free of obstacles. This is, in particular, the case if the robot operates in a dynamic environment, for example in one used by humans. In dynamic, real-world environments the robot has to face situations in which doors are closed, that there are obstacles on the planned path or that the environment has changed and the given map is, therefore, no longer valid. One of the most popular approaches to attack this problem is to equip the robot with sensors that allow it to measure the distance to obstacles and thus avoid obstacles. Additionally, an approach is used that avoids collisions and makes dynamic adjustments to the previously planned path. In order to navigate along a planned path, the robot must actually be able to accurately determine its position on the map and on the planned path (or distance from it). For this purpose, current navigation systems for robots use special algorithms based on probabilistic principles,[22] such as the *Kalman* filter[23] or the particle filter algorithm.[24] Both approaches and their variants have been shown to be extremely robust for determining a probability distribution about the position of the vehicle based on the distances to obstacles determined by the distance sensor and the given obstacle map. Given this distribution, the robot can choose its most likely position to make its navigation decisions. The majority of autonomously navigating robots that are not guided by induction loops, optical markers, or lines utilize probabilistic approaches for robot localization. A basic requirement for the components discussed thus far is the existence of a map. But how can a robot obtain such an obstacle map? In principle, there are two possible solutions for this. First, the user can measure the environment and use it to create a map with the exact positions of all objects in the robot's workspace. This map can then be used to calculate the position of the vehicle or to calculate paths in the environment. The alternative is to use a so-called SLAM (Simultaneous Localization and Mapping)[25] method. Here, the robot is steered through its environment and, based on the data gathered throughout this process, automatically computes the map. Incidentally, this SLAM

[21] Hart and others, 'A formal basis for the heuristic determination of minimum cost paths' (n 8).
[22] S Thrun, W Burgard, and D Fox, *Probabilistic Robotics* (2005) (hereafter Thrun and others, *Probabilistic Robotics*).
[23] RE Kalman, 'A New Approach to Linear Filtering and Prediction Problems' (1960) *ASME–Journal of Basic Engineering* 35.
[24] D Fox and others, 'Monte Carlo Localization: Efficient Position Estimation for Mobile Robots' (1999) Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI) 343.
[25] Thrun and others, *Probabilistic Robotics* (n 22).

technique is also known in photogrammetry where it is used for generating maps based on measurements.[26] These four components: path planning, collision avoidance and replanning, localization, and SLAM for map generation are key to today's navigation robots and also self-driving cars.

2. Key Technology Autonomous Manipulation

Manipulation has been successfully used in production processes in the past. The majority of these robots had fixed programmed actions and, furthermore, a cage around them to prevent humans from entering the action spaces of the robots. The future, however, lies in robots that are able to robustly grasp arbitrary objects even from cluttered scenes and that are intrinsically safe and cannot harm people. In particular, the development of lightweight systems[27] will be a key enabler for human–robot collaboration. On the other hand, this requires novel approaches to robust manipulation. In this context, again, AI technology based on deep learning has played a key role over the past years and is envisioned to provide innovative solutions for the future. Recently, researchers presented an approach to apply deep learning to robustly grasp objects from cluttered scenes.[28] Both approaches will enable us in the future to build robots that coexist with humans, learn from them, and improve over time.

IV. CURRENT AND FUTURE FIELDS OF APPLICATION AND CHALLENGES

As already indicated, AI is currently more and more becoming a part of our daily lives. This affects both our personal and professional lives. Important transporters of AI technology are smartphones, as numerous functions on them are based on AI. For example, we can already control them by voice, they recognize faces in pictures, they automatically store important information for us, such as where our car is parked, and they play music we like after analyzing our music library or learning what we like from our ratings of music tracks. By analyzing these preferences in conjunction with those of other users, the predictions of tracks we like get better and better. This can, of course, be applied to other activities, such as shopping, where shopping platforms suggest possible products we might be interested in. This has long been known from search engines, which try to present us with answers that correspond as closely as possible to the Web pages for which we are actually looking. In robotics, the current key areas are logistics and flexible production (Industry 4.0). To remain competitive, companies must continue to optimize production processes. Here, mobile robots and flexible manipulation systems that can cooperate with humans will play a decisive role. This will result in significantly more flexible production processes, which will be of enormous importance for all countries with large manufacturing sectors. However, robots are also envisioned to perform various tasks in our homes.

By 2030, AI will penetrate further areas: Not only will we see robots performing ever more demanding tasks in production, but also AI techniques will find their way into areas performed

[26] P Agarwal, W Burgard, and C Stachniss, 'Survey of Geodetic Mapping Methods: Geodetic Approaches to Mapping and the Relationship to Graph-Based SLAM' (2014) 21(3) *IEEE Robotics & Automation Magazine* 63.

[27] G Hirzinger and others, 'On a New Generation of Torque Controlled Light-Weight Robots' (2001) 4 *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* 3356.

[28] J Mahler and others, 'Dex-net 1.0: A Cloud-Based Network of 3d Objects for Robust Grasp Planning Using a Multi-Armed Bandit Model with Correlated Rewards' (2016) Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) 1957.

by people with highly qualified training. For example, there was a paper in *Nature* that presented a system that could diagnose skin cancer based on an image of the skin taken with a cell phone.[29] The interesting aspect of this work is that the authors were actually able to achieve the detection rate of dermatologists with their deep neural networks-based system. This clearly indicates that there is enormous potential in AI to further optimize processes that require a high level of expertise.

With the increasing number of applications of systems relying on AI technology, there is also a growing need for the responsibility or the responsible governance of such systems. In particular, when they can impose risks for individuals, for example in the context of service robots that collaborate with humans or self-driving cars that co-exist with human traffic participants, where mistakes of the physical agent might substantially harm a person, the demands for systems whose behavior can be explained to, or understood by, humans are high. Even in the context of risk-free applications, there can be such a demand, for example, to better identify biases in recommender systems. A further relevant issue is that of privacy. In particular, AI systems based on machine learning require a large amount of data, which imposes the question of how these systems can be trained so that the privacy of the users can be maintained while at the same time providing all the necessary benefits. A further interesting tool for advancing the capabilities of such systems is fleet learning, learning in which all systems jointly learn from their users how to perform specific tasks. In this context, the question arises of how to guarantee that no system is taught inappropriate or even dangerous behavior. How can we build such systems so that they conform with values, norms, and regulations? Answers to these questions are by themselves challenging research problems and many chapters in this book address them.

[29] A Esteva and others, 'Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks' (2017) 542(7639) *Nature* 115.

2

Automating Supervision of AI Delegates

Jaan Tallinn and Richard Ngo

As the field of machine learning advances, AI systems are becoming more and more useful in a number of domains, in particular due to their increasing ability to generalise beyond their training data. Our focus in this chapter is on understanding the different possibilities for the deployment of highly capable and general systems which we may build in the future. We introduce a framework for the deployment of AI which focuses on two ways for humans to interact with AI systems: delegation and supervision. This framework provides a new lens through which to view both the relationship between humans and AIs, and the relationship between the training and deployment of machine learning systems.

1. AIS AS TOOLS, AGENTS, OR DELEGATES

The last decade has seen dramatic progress in Artificial Intelligence (AI), in particular due to advances in deep learning and reinforcement learning. The increasingly impactful capabilities of our AI systems raise important questions about what future AIs might look like and how we might interact with them. In one sense, AI can be considered a particularly sophisticated type of software. Indeed, the line between AI and other software is very blurry: many software products rely on algorithms which fell under the remit of AI when they were developed, but are no longer typically described as AI.[1] Prominent examples include search engines like Google and image-processing tools like optical character recognition. Thus, when thinking about future AI systems, one natural approach is to picture us interacting with them similarly to how we interact with software programs: as tools which we will use to perform specific tasks, based on predefined affordances, via specially-designed interfaces.

Let us call this the 'tool paradigm' for AI. Although we will undoubtedly continue to develop some AIs which fit under this paradigm, compelling arguments have been made that other AIs will fall outside it – in particular, AIs able to flexibly interact with the real world to perform a wide range of tasks, displaying general rather than narrow intelligence. The example of humans shows that cognitive skills gained in one domain can be useful in a wide range of other domains; it is difficult to argue that the same cannot be true for AIs, especially given the similarities between human brains and deep neural networks. Although no generally intelligent AIs exist today, and some AI researchers are skeptical about the prospects for building them, most expect

[1] M Minsky, 'Thoughts about Artificial Intelligence' in R Kurzweil (ed), *The Age of Intelligent Machines* (1990).

it to be possible within this century.[2] However, it does not require particularly confident views on the timelines involved to see value in starting to prepare for the development of artificial general intelligence (AGI) already.

Why won't AGIs fit naturally into the tool paradigm? There are two core reasons: flexibility and autonomy. Tools are built with a certain set of affordances, which allow a user to perform specific tasks with them.[3] For example, software programs provide interfaces for humans to interact with, where different elements of the interface correspond to different functionalities. However, predefined interfaces cannot adequately capture the wide range of tasks that humans are, and AGIs will be, capable of performing. When working with other humans, we solve this problem by using natural language to specify tasks in an expressive and flexible way; we should expect that these and other useful properties will ensure that natural language is a key means of interacting with AGIs. Indeed, AI assistants such as Siri and Alexa are already rapidly moving in this direction.

A second difference between using tools and working with humans: when we ask a human to perform a complex task for us, we don't need to directly specify each possible course of action. Instead, they will often be able to make a range of decisions and react to changing circumstances based on their own judgements. We should expect that, in order to carry out complex tasks like running a company, AGIs will also need to be able to act autonomously over significant periods of time. In such cases, it seems inaccurate to describe them as tools being directly used by humans, because the humans involved may know very little about the specific actions the AGI is taking.

In an extreme case, we can imagine AGIs which possess ingrained goals which they pursue autonomously over arbitrary lengths of time. Let's call this the full autonomy paradigm. Such systems have been discussed extensively by *Nick Bostrom* and *Eliezer Yudkowsky*.[4] *Stuart Russell* argues that they are the logical conclusion of extrapolating the current aims and methods of machine learning.[5] Under this paradigm, AIs would acquire goals during their training process which they then pursue throughout deployment. Those goals might be related to, or influenced by, human preferences and values, but could be pursued without humans necessarily being in control or having veto power.

The prospect of creating another type of entity which independently pursues goals in a similar way to humans raises a host of moral, legal, and safety questions, and may have irreversible effects – because once created, autonomous AIs with broad goals will have incentives to influence human decision-making towards outcomes more favourable to their goals. In particular, concerns have been raised about the difficulty of ensuring that goals acquired by AIs during training are desirable ones from a human perspective. Why might AGIs nevertheless be built with this level of autonomy? The main argument towards this conclusion is that increasing AI autonomy will be a source of competitive economic or political advantage, especially if an AGI race occurs.[6] Once an AI's strategic decision-making abilities exceed those of humans, then the ability to operate independently, without needing to consult humans and wait for their decisions, would

[2] K Grace and others, 'When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 *Journal of Artificial Intelligence Research* 729.

[3] JJ Gibson, 'The Theory of Affordances' in JJ Gibson (ed), *The Ecological Approach to Visual Perception* (1979) 127–137.

[4] N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014); E Yudkowsky, 'Artificial Intelligence as a Positive and Negative Factor in Global Risk' in N Bostrom and MM Cirkovic (eds), *Global Catastrophic Risks* (2008) 184.

[5] S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

[6] S Cave and S ÓhÉigeartaigh, 'An AI Race for Strategic Advantage: Rhetoric and Risks' in J Furman and others (eds), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) 36–40.

give it a speed advantage over more closely-supervised competitors. This phenomenon has already been observed in high-frequency trading in financial markets – albeit to a limited extent, because trading algorithms can only carry out a narrow range of predefined actions.

Authors who have raised these concerns have primarily suggested that they be solved by developing better techniques for building the *right* autonomous AIs. However, we should not consider it a foregone conclusion that we will build fully autonomous AIs at all. As *Stephen Cave* and *Sean ÓhÉigeartaigh* point out, AI races are driven in part by self-fulfilling narratives – meaning that one way of reducing their likelihood is to provide alternative narratives which don't involve a race to fully autonomous AGI.[7] In this chapter we highlight and explore an alternative which lies between the tool paradigm and the full autonomy paradigm, which we call the supervised delegation paradigm. The core idea is that we should aim to build AIs which can perform tasks and make decisions on our behalf upon request, but which lack persistent goals of their own outside the scope of explicit delegation. Like autonomous AIs, delegate AIs would be able to infer human beliefs and preferences, then flexibly make and implement decisions without human intervention; but like tool AIs, they would lack agency when they have not been deployed by humans. We call systems whose motivations function in this way aligned delegates (as discussed further in the next section).

The concept of delegation has appeared in discussions of agent-based systems going back decades,[8] and is closely related to *Bostrom's* concept of 'genie AI'.[9] Another related concept is the AI assistance paradigm advocated by Stuart Russell, which also focuses on building AIs that pursue human goals rather than their own goals.[10] However, Russell's conception of assistant AIs is much broader in scope than delegate AIs as we have defined them, as we discuss in the next section. More recently, delegation was a core element of *Andrew Critch* and *David Krueger's* ARCHES framework, which highlights the importance of helping multiple humans safely delegate tasks to multiple AIs.[11]

While most of the preceding works were motivated by concern about the difficulty of alignment, they spend relatively little time explaining the specific problems involved in aligning machine learning systems, and how proposed solutions address them. The main contribution of this chapter is to provide a clearer statement of the properties which we should aim to build into AI delegates, the challenges which we should expect, and the techniques which might allow us to overcome them, in the context of modern machine learning (and more specifically deep reinforcement learning). A particular focus is the importance of having large amounts of data which specify desirable behaviour – or, in more poetic terms, the 'unreasonable effectiveness of data'.[12] This is where the supervised aspect of supervised delegation comes in: we argue that, in order for AI delegates to remain trustworthy, it will be necessary to continuously monitor and evaluate their behaviour. We discuss ways in which the difficulties of doing so give rise to a tradeoff between safety and autonomy. We conclude with a discussion of how the goal of alignment can be a focal point for cooperation, rather than competition, between groups involved with AI development.

[7] Ibid.

[8] C Castelfranchi and R Falcone, 'Towards a Theory of Delegation for Agent-Based Systems' (1998) 24(3–4) *Robotics and Autonomous systems* 141.

[9] N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014).

[10] S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

[11] A Critch and D Krueger, 'AI Research Considerations for Human Existential Safety (ARCHES)' (*arXiv*, 30 May 2020) https://arxiv.org/abs/2006.04948v1.

[12] A Halevy, P Norvig, and F Pereira, 'The Unreasonable Effectiveness of Data' (2009) 24(2) *IEEE Intelligent Systems*, 8–12.

II. ALIGNED DELEGATES

What does it mean for an AI to be aligned with a human? The definition which we will use here comes from *Paul Christiano*: an AI is intent aligned with a human if the AI is trying to do what the human wants it to do.[13] To be clear, this does not require that the AI is correct about what the human wants it to do, nor that it succeeds – both of which will be affected by the difficulty of the task and the AI's capabilities. The concept of intent alignment (henceforth just 'alignment') instead attempts to describe an AI's motivations in a way that's largely separable from its capabilities.

Having said that, the definition still assumes a certain baseline level of capabilities. As defined above, alignment is a property only applicable to AIs with sufficiently sophisticated motivational systems that they can be accurately described as trying to achieve things. It also requires that they possess sufficiently advanced theories of mind to be able to ascribe desires and intentions to humans, and reasonable levels of coherence over time. In practice, because so much of human communication happens via natural language, it will also require sufficient language skills to infer humans' intentions from their speech. Opinions differ on how difficult it is to meet these criteria – some consider it appropriate to take an 'intentional stance' towards a wide range of systems, including simple animals, whereas others have more stringent requirements for ascribing intentionality and theory of mind.[14] We need not take a position on these debates, except to hold that sufficiently advanced AGIs could meet each of these criteria.

Another complication comes from the ambiguity of 'what the human wants'. *Iason Gabriel* argues that 'there are significant differences between AI that aligns with instructions, intentions, revealed preferences, ideal preferences, interests and values'; *Christiano's* definition of alignment doesn't pin down which of these we should focus on.[15] Alignment with the ideal preferences and values of fully-informed versions of ourselves (also known as 'ambitious alignment') has been the primary approach discussed in the context of fully autonomous AI. Even *Russell's* assistant AIs are intended to 'maximise the realisation of human preferences' – where he is specifically referring to preferences that are 'all-encompassing: they cover everything you might care about, arbitrarily far into the future'.[16]

Yet it's not clear whether this level of ambitious alignment is either feasible or desirable. In terms of feasibility, focusing on long timeframes exacerbates many of the problems we discuss in later sections. And in terms of desirability, ambitious alignment implies that a human is no longer an authoritative source for what an AI aligned with that human should aim to do. An AI aligned with a human's revealed preferences, ideal preferences, interests, or values might believe that it understands them better than the human does, which could lead to that AI hiding information from the human or disobeying explicit instructions. Because we are still very far from any holistic theory of human preferences or values, we should be wary of attempts to design AIs which take actions even when their human principals explicitly instruct them not to; let us call this the principle of deference. (Note that the principle is formulated in an asymmetric way – it seems plausible that aligned AIs should sometimes avoid taking actions even when instructed to do so, in particular illegal or unethical actions.)

[13] P Christiano, 'Clarifying "AI Alignment"' (*AI Alignment*, 7 April 2018) https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6.
[14] DC Dennett, 'Précis of the Intentional Stance' (1988) 11 *Behavioral and Brain Stances* 495.
[15] I Gabriel, 'Artificial Intelligence, Values, and Alignment' (2020) 30 *Minds and Machines* 411.
[16] S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019).

For our purposes, then, we shall define a delegate AI as aligned with a human principal if it tries to do only what that human intends it to do, where the human's intentions are interpreted to be within the scope of tasks the AI has been delegated. What counts as a delegated task depends on what the principal has said to the AI – making natural language an essential element of the supervised delegation paradigm. This contrasts with both the tool paradigm (in which many AIs will not be general enough to understand linguistic instructions) and the full autonomy paradigm (in which language is merely considered one of many information channels which help AIs understand how to pursue their underlying goals).

Defining delegation in terms of speech acts does not, however, imply that all relevant information needs to be stated explicitly. Although early philosophers of language focused heavily on the explicit content of language, more recent approaches have emphasised the importance of a pragmatic focus on speaker intentions and wider context in addition to the literal meanings of the words spoken.[17] From a pragmatic perspective, full linguistic competence includes the ability to understand the (unspoken) implications of a statement, as well as the ability to interpret imprecise or metaphorical claims in the intended way. Aligned AI delegates should use this type of language understanding in order to interpret the 'scope' of tasks in terms of the pragmatics and context of the instructions given, in the same way that humans do when following instructions. An AI with goals which extend outside that scope, or which don't match its instructions, would count as misaligned.

We should be clear that aiming to build aligned delegates with the properties described above will likely involve making some tradeoffs against desirable aspects of autonomy. For example, an aligned delegate would not take actions which are beneficial for its user that are outside the scope of what it has been asked to do; nor will it actively prevent its user from making a series of bad choices. We consider these to be features, though, rather than bugs – they allow us to draw a boundary before reaching full autonomy, with the aim of preventing a gradual slide into building fully autonomous systems before we have a thorough understanding of the costs, benefits, and risks involved. The clearer such boundaries are, the easier it will be to train AIs with corresponding motivations (as we discuss in the next section).

A final (but crucial) consideration is that alignment is a two-place predicate: an AI cannot just be aligned simpliciter, but rather must be aligned with a particular principal – and indeed could be aligned with different principals in different ways. For instance, when AI developers construct an AI, they would like it to obey the instructions given to it by the end user, but only within the scope of whatever terms and conditions have been placed on it. From the perspective of a government, another limitation is desirable: AI should ideally be aligned to their end users only within the scope of legal behaviour. The questions of who AIs should be aligned with, and who should be held responsible for their behaviour, are fundamentally questions of politics and governance rather than technical questions. However, technical advances will affect the landscape of possibilities in important ways. Particularly noteworthy is the effect of AI delegates performing impactful political tasks – such as negotiation, advocacy, or delegation of their own – on behalf of their human principals. The increased complexity of resulting AI governance problems may place stricter requirements on technical approaches to achieving alignment.[18]

[17] K Korta and J Perry, 'Pragmatics' (*The Stanford Encyclopedia of Philosophy*, 21 August 2019) https://plato.stanford.edu/archives/fall2019/entries/pragmatics/.

[18] A Critch and D Krueger, 'AI Research Considerations for Human Existential Safety (ARCHES)' (*arXiv*, 30 May 2020) https://arxiv.org/abs/2006.04948v1.

III. THE NECESSITY OF HUMAN SUPERVISION

So far we have talked about desirable properties of alignment without any consideration of how to achieve those desiderata. Unfortunately, a growing number of researchers have raised concerns that current machine learning techniques are inadequate for ensuring alignment of AGIs. Research in this area focuses on two core problems. The first is the problem of outer alignment: the difficulty in designing reward functions for reinforcement learning agents which incentivise desirable behaviour while penalising undesirable behaviour.[19] *Victoria Krakovna* et al catalogue many examples of specification gaming in which agents find unexpected ways to score highly even in relatively simple environments, most due to mistakes in how the reward function was specified.[20] As we train agents in increasingly complex and open-ended environments, designing ungameable reward functions will become much more difficult.[21]

One major approach to addressing the problems with explicit reward functions involves generating rewards based on human data – known as reward learning. Early work on reward learning focused on deriving reward functions from human demonstrations – a process known as inverse reinforcement learning.[22] However, this requires humans themselves to be able to perform the task to a reasonable level in order to provide demonstrations. An alternative approach which avoids this limitation involves inferring reward functions from human evaluations of AI behaviour. This approach, known as reward modelling, has been used to train AIs to perform tasks which humans cannot demonstrate well, such as controlling a (simulated) robot body to do a backflip.[23]

In most existing examples of reward learning, reward functions are learned individually for each task of interest – an approach which doesn't scale to systems which generalise to new tasks after deployment. However, a growing body of work on interacting with reinforcement learning agents using natural language has been increasingly successful in training AIs to generalise to novel instructions.[24] This fits well with the vision of aligned delegation described in the previous section, in which specification of tasks for AIs involves two steps: first training AIs to have aligned motivations, and then using verbal instructions to delegate them specific tasks. The hope is that if AIs are rewarded for following a wide range of instructions in a wide range of situations, then they will naturally acquire the motivation to follow human instructions in general, including novel instructions in novel environments.

However, this hope is challenged by a second concern. The problem of inner alignment is that even if we correctly specify the reward function used during training, the resulting policy may not possess the goal described by that reward function. In particular, it may learn to pursue proxy goals which are correlated with reward during most of the training period, but which eventually diverge (either during later stages of training, or during deployment).[25] This possibility is analogous to how humans learned to care directly about food, survival, sex, and so

[19] Alignment problems also exist for AIs trained in other ways, such as self-supervised learning; here I focus on the case of reinforcement learning for the sake of clarity.

[20] V Krakovna and others, 'Specification Gaming: The Flip Side of AI Ingenuity' (*Deep Mind*, 2020) deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

[21] A Ecoffet, J Clune, and J Lehman, 'Open Questions in Creating Safe Open-Ended AI: Tensions between Control and Creativity' in *Artificial Life Conference Proceedings* (2020) 27–35.

[22] AY Ng and SJ Russell, 'Algorithms for Inverse Reinforcement Learning' (2000) in 1 Icml 2.

[23] P Christiano and others, 'Deep Reinforcement Learning from Human Preferences' (*arXiv*, 13 July 2017).

[24] J Luketina and others, 'A Survey of Reinforcement Learning Informed by Natural Language' (*arXiv*,10 June 2019) https://arxiv.org/abs/1906.03926; J Abramson and others, 'Imitating Interactive Intelligence' (*arXiv*, 21 January 2021).

[25] J Koch and others, 'Objective Robustness in Deep Reinforcement Learning'(*arXiv*, 8 June 2021).

on – proxies which were strongly correlated with genetic fitness in our ancestral environment, but are much less so today.[26] As an illustration of how inner misalignment might arise in the context of machine learning, consider training a policy to follow human instructions in a virtual environment containing many incapacitating traps. If it is rewarded every time it successfully follows an instruction, then it will learn to avoid becoming incapacitated, as that usually prevents it from completing its assigned task. This is consistent with the policy being aligned, if the policy only cares about surviving the traps as a means to complete its assigned task – in other words, as an instrumental goal. However, if policies which care about survival only as an instrumental goal receive (nearly) the same reward as policies which care about survival for its own sake (as a final goal) then we cannot guarantee that the training process will find one in the former category rather than the latter.

Now, survival is just one example of a proxy goal that might lead to inner misalignment; and it will not be relevant in all training environments. But training environments which are sufficiently complex to give rise to AGI will need to capture at least some of the challenges of the real world – imperfect information, resource limitations, and so on. If solving these challenges is highly correlated with receiving rewards during training, then how can we ensure that policies only learn to care about solving those challenges for instrumental purposes, within the bounds of delegated tasks? The most straightforward approach is to broaden the range of training data used, thereby reducing the correlations between proxy goals and the intended goal. For example, in the environment discussed in the previous paragraph, instructing policies to deliberately walk into traps (and rewarding them for doing so) would make survival less correlated with reward, thereby penalising policies which pursue survival for its own sake.

In practice, though, when talking about training artificial general intelligences to perform a wide range of tasks, we should expect that the training data will encode many incentives which are hard to anticipate in advance. Language models such as GPT-3 are already being used in a wide range of surprising applications, from playing chess (using text interactions only) to generating text adventure games.[27] It will be difficult for AI developers to monitor AI behaviour across many domains, and then design rewards which steer those AIs towards intended behaviour. This difficulty is exacerbated by the fact that modern machine learning techniques are incredibly data-hungry: training agents to perform well in difficult games can take billions of steps. If the default data sources available give rise to inner or outer alignment problems, then the amount of additional supervision required to correct these problems may be infeasible for developers to collect directly. So, how can we obtain enough data to usefully think about alignment failures in a wide range of circumstances, to address the outer and inner alignment problems?

Our suggestion is that this gap in supervision can be filled by end users. Instead of thinking of AI development as a training process followed by a deployment process, we should think of it as an ongoing cycle in which users feed back evaluations which are then used to help align future AIs. In its simplest form, this might involve users identifying inconsistencies or omissions in an AI's statements, or ways in which it misunderstood the user's intentions, or even just occasions when it took actions without having received human instructions. In order to further constrain an AI's autonomy, the AI can also be penalised for behaviour which was desirable, but beyond

[26] E Hubinger and others, 'Risks from Learned Optimization in Advanced Machine Learning Systems' (*arXiv*, 11 June 2019).

[27] S Alexander, 'A Very Unlikely Chess Game' (*Slate Star Codex*, 6 January 2020) https://slatestarcodex.com/2020/01/06/a-very-unlikely-chess-game/;N Walton, 'AI Dungeon: Dragon Model Upgrade' (*Latitude Team*, 14 July 2020) https://aidungeon.medium.com/ai-dungeon-dragon-model-upgrade-7e8ea579abfe.

the scope of the task it was delegated to perform. This form of evaluation is much easier than trying to evaluate the long-term consequences of an AI's actions; yet it still pushes back against the underlying pressure towards convergent instrumental goals and greater autonomy that we described above.

Of course, user data is already collected by many different groups for many different purposes. Prominent examples include scraping text from Reddit, or videos from YouTube, in order to train large self-supervised machine learning models. However, these corpora contain many examples of behaviour we wouldn't like AIs to imitate – as seen in GPT-3's regurgitation of stereotypes and biases found in its training data.[28] In other cases, evaluations are inferred from user behaviour: likes on a social media post, or clicks on a search result, can be interpreted as positive feedback. Yet these types of metrics already have serious limitations: there are many motivations driving user engagement, not all of which should be interpreted as positive feedback. As interactions with AI become much more freeform and wide-ranging, inferred correlations will become even less reliable, compared with asking users to evaluate AI alignment directly. So even if users only perform explicit evaluations of a small fraction of AI behaviour, this could provide much more information about their alignment than any other sources of data currently available. And, unlike other data sources, user evaluations could flexibly match the distributions of tasks on which AIs are actually deployed in the real world, and respond to new AI behaviour very quickly.[29]

IV. BEYOND HUMAN SUPERVISION

Unfortunately, there are a number of reasons to expect that even widespread use of human evaluation will not be sufficient for reliable supervision in the long term. The core problem is that the more sophisticated an AI's capabilities are, the harder it is to identify whether it is behaving as intended or not. In some narrow domains like chess and Go, experts already struggle to evaluate the quality of AI moves, and to tell the difference between blunders and strokes of brilliance. The much greater complexity of the real world will make it even harder to identify all the consequences of decisions made by AIs, especially in domains where they make decisions far faster and generate much more data than humans can keep up with.

Particularly worrying is the possibility of AIs developing deceptive behaviour with the aim of manipulating humans into giving better feedback. The most notable example of this came from reward modelling experiments in which a human rewarded an AI for grasping a ball with a robotic claw.[30] Instead of completing the intended task, the AI learned to move the claw into a position between the camera and the ball, thus appearing to grasp the ball without the difficulty of actually doing so. As AIs develop a better understanding of human psychology and the real-world context in which they're being trained, manipulative strategies like this could become much more complex and much harder to detect. They would also not necessarily be limited to affecting observations sent directly to humans, but might also attempt to modify their reward signal using any other mechanisms they can gain access to.

[28] TB Brown and others, 'Language Models Are Few-Shot Learners' (*arXiv*, 22 July 2020) https://arxiv.org/abs/2005.14165?source=techstories.org.

[29] This does assume a high level of buy-in from potential users, which may be difficult to obtain given privacy concerns. We hope that the project of alignment can be presented in a way that allows widespread collaboration – as discussed further in the final section of this chapter.

[30] P Christiano and others, 'Deep Reinforcement Learning from Human Preferences' (*arXiv*, 13 July 2017) https://arxiv.org/abs/1706.03741.

The possibility of manipulation is not an incidental problem, but rather a core difficulty baked into the use of reinforcement learning in the real world. As AI pioneer *Stuart Russell* puts it:

> The formal model of reinforcement learning assumes that the reward signal reaches the agent from outside the environment; but [in fact] the human and robot are part of the same environment, and the robot can maximize its reward by modifying the human to provide a maximal reward signal at all times. . . . [This] indicates a fundamental flaw in the standard formulation of RL.[31]

In other words, AIs are trained to score well on their reward functions by taking actions to influence the environment around them, and human supervisors are a part of their environment which has a significant effect on the reward they receive, so we should expect that by default AIs will learn to influence their human supervisors. This can be mitigated if supervisors heavily penalise attempted manipulation when they spot it – but this still leaves an incentive for manipulation which can't be easily detected. As AIs come to surpass human abilities on complex real-world tasks, preventing them from learning manipulative strategies will become increasingly difficult – especially if AI capabilities advance rapidly, so that users and researchers have little time to notice and respond to the problem.

How might we prevent this, if detecting manipulation or other undesirable behaviour eventually requires a higher quality and quantity of evaluation data than unaided humans can produce? The main mechanisms which have been proposed for expanding the quality/quantity frontier of supervision involve relying on AI systems themselves to help us supervise other AIs. When considering this possibility, we can reuse two of the categories discussed in the first section: we can either have AI-based supervision tools, or else we can delegate the process of supervision to another AI (which we shall call recursive supervision, as it involves an AI delegate supervising another AI delegate, which might then supervise another AI delegate, which. . .).

One example of an AI-based supervision tool is a reward model which learns to imitate human evaluations of AI behaviour. Reinforcement learning agents whose training is too lengthy for humans to supervise directly (e.g. involving billions of steps) can then be supervised primarily by reward models instead. Early work on reward models demonstrated a surprising level of data efficiency: reward models can greatly amplify a given amount of human feedback.[32] However, the results of these experiments also highlighted the importance of continual feedback – when humans stopped providing new data, agents eventually found undesirable behaviours which nevertheless made the reward models output high scores.[33] So reward models are likely to rely on humans continually evaluating AI behaviour as it expands into new domains.

Another important category of supervision tool is interpretability tools, which aim to explain the mechanisms by which a system decides how to act. Although deep neural networks are generally very opaque to mechanistic explanation, there has been significant progress over the last few years in identifying how groups of artificial neurons (and even individual neurons) contribute to the overall output.[34] One long-term goal of this research is to ensure that AIs will honestly explain the reasoning that led to their actions and their future intentions. This would help address the inner alignment problems described above, because agents could be penalised

[31] S Russell, 'Provably Beneficial Artificial Intelligence' in A de Grey and others (eds), *Exponential Life, The Next Step* (2017).

[32] P Christiano and others, 'Deep Reinforcement Learning from Human Preferences' (*arXiv*, 13 July 2017) https://arxiv.org/abs/1706.03741.

[33] B Ibarz and others, 'Reward Learning from Human Preferences and Demonstrations in Atari' (*arXiv*, 15 November 2018) https://arxiv.org/abs/1811.06521.

[34] N Cammarata and others, 'Thread: Circuits' (*Distill*, 10 March 2020) https://distill.pub/2020/circuits/.

for acting according to undesirable motivations even when their behaviour is indistinguishable from the intended behaviour. However, existing techniques are still far from being able to identify deceptiveness (or other comparably abstract traits) in sophisticated models.

Recursive supervision is currently also in a speculative position, but some promising strategies have been identified. A notable example is *Geoffrey Irving, Paul Christiano,* and *Dario Amodei's* Debate technique, in which two AIs are trained to give arguments for opposing conclusions, with a human judging which arguments are more persuasive.[35] Because the rewards given for winning the debate are zero-sum, the resulting competitive dynamic should in theory lead each AI to converge towards presenting compelling arguments which are hard to rebut – analogous to how AIs trained via self-play on zero-sum games converge to winning strategies. However, two bottlenecks exist: the ease with which debaters can identify flaws in deceptive arguments, and the accuracy with which humans can judge allegations of deception. Several strategies have been proposed to make judging easier – for example, incorporating cross-examination of debaters, or real-world tests of claims made during the debate – but much remains to be done in fleshing out and testing Debate and other forms of recursive supervision.

To some extent, recursive supervision will also arise naturally when multiple AIs are deployed in real-world scenarios. For example, if one self-driving car is driving erratically, then it's useful for others around it to notice and track that. Similarly, if one trading AI is taking extreme positions that move the market considerably, then it's natural for other trading AIs to try to identify what's happening and why. This information could just be used to flag the culprit for further investigation – but it could also be used as a supervision signal for further training, if shared with the relevant AI developers. In the next section we discuss the incentives which might lead different groups to share such information, or to cooperate in other ways.

V. AI SUPERVISION AS A COOPERATIVE ENDEAVOUR

We started this chapter by discussing some of the competitive dynamics which might be involved in AGI development. However, there is reason to hope that the process of increasing AI alignment is much more cooperative than the process of increasing AI capabilities. This is because misalignment could give rise to major negative externalities, especially if misaligned AIs are able to accumulate significant political, economic, or technological power (all of which are convergent instrumental goals). While we might think that it will be easy to 'pull the plug' on misbehaviour, this intuition fails to account for strategies which highly capable AIs might use to prevent us from doing so – especially those available to them after they have already amassed significant power. Indeed, the history of corporations showcases a range of ways that 'agents' with large-scale goals and economic power can evade oversight from the rest of society. And AIs might have much greater advantages than corporations currently do in avoiding accountability – for example, if they operate at speeds too fast for humans to monitor. One particularly stark example of how rapidly AI behaviour can spiral out of control was the 2010 Flash Crash, in which high-frequency trading algorithms got into a positive feedback loop and sent prices crashing within a matter of minutes.[36] Although the algorithms involved were relatively simple by the standards of modern machine learning (making this an example of accidental failure rather than misalignment),

[35] G Irving, P Christiano, and D Amodei, 'AI Safety via Debate' (*arXiv*, 22 October 2018) https://arxiv.org/abs/1805.00899.

[36] US Securities & Exchange Commission and US Commodity Futures Trading Commission, 'Findings Regarding the Market Events of May 6, 2010. Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues' (*US Securities and Exchange Commission*, 30 September 2010) www.sec.gov/news/studies/2010/marketevents-report.pdf.

AIs sophisticated enough to reason about the wider world will be able to deliberately implement fraudulent or risky behaviour at increasingly bewildering scales.

Preventing them from doing so is in the interests of humanity as a whole; but what might large-scale cooperation to improve alignment actually look like? One possibility involves the sharing of important data – in particular data which is mainly helpful for increasing AI's alignment rather than their capabilities. It is somewhat difficult to think about how this would work for current systems, as they don't have the capabilities identified as prerequisites for being aligned in section 2. But as one intuition pump for how sharing data can differentially promote safety over other capabilities, consider the case of self-driving cars. The data collected by those cars during deployment is one of the main sources of competitive advantage for the companies racing towards autonomous driving, making them rush to get cars on the road. Yet, of that data, only a tiny fraction consists of cases where humans are forced to manually override the car steering, or where the car crashes. So while it would be unreasonably anticompetitive to force self-driving car companies to share all their data, it seems likely that there is some level of disclosure which contributes to preventing serious failures much more than to erasing other competitive advantages. This data could be presented in the form of safety benchmarks, simple prototypes of which include DeepMind's AI Safety Gridworlds and the Partnership on AI's SafeLife environment.[37]

The example of self-driving cars also highlights another factor which could make an important contribution to alignment research: increased cooperation amongst researchers thinking about potential risks from AI. There is currently a notable divide between researchers primarily concerned about near-term risks and those primarily concerned about long-term risks.[38] Currently, the former tend to focus on supervising the activity of existing systems, whereas the latter prioritise automating the supervision of future systems advanced enough to be qualitatively different from existing systems. But in order to understand how to supervise future AI systems, it will be valuable to have access not only to technical research on scalable supervision techniques, but also to hands-on experience of how supervision of AIs works in real-world contexts and the best practices identified so far. So, as technologies like self-driving cars become increasingly important, we hope that the lessons learned from their deployment can help inform work on long-term risks via collaboration between the two camps.

A third type of cooperation to further alignment involves slowing down capabilities research to allow more time for alignment research to occur. This would require either significant trust between the different parties involved, or else strong enforcement mechanisms.[39] However, cooperation can be made easier in a number of ways. For example, corporations can make themselves more trustworthy via legal commitments such as windfall clauses.[40] A version of this has already been implemented in OpenAI's capped-profit structure, along with other innovative legal mechanisms – most notably the clause in OpenAI's charter which commits to assisting rather than competing with other projects, if they meet certain conditions.[41]

[37] J Leike and others, 'AI Safety Gridworlds' (*arXiv*, 28 November 2017) https://arxiv.org/abs/1711.09883; CL Wainwright and P Eckersley, 'Safelife 1.0: Exploring Side Effects in Complex Environments' (*arXiv*, 26 February 2021) https://arxiv.org/abs/1912.01217.

[38] S Cave and S ÓhÉigeartaigh, 'Bridging Near-and Long-Term Concerns about AI' (2019) 1 *Nature Machine Intelligence* 5–6.

[39] A Dafoe, 'AI Governance: A Research Agenda' in *Governance of AI Program*, Future of Humanity Institute, 2017 www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf, 1–53.

[40] C O'Keefe and others, 'The Windfall Clause: Distributing the Benefits of AI for the Common Good' in J Furman and others, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018) 327–331.

[41] OpenAI, 'OpenAI Charter' (9 April 2018) https://openai.com/charter/; OpenAI, 'OpenAI LP' (11 March 2019) https://openai.com/blog/openai-lp/.

We are aware that we have skipped over many of the details required to practically implement large-scale cooperation to increase AI alignment – some of which might not be pinned down for decades to come. Yet we consider it important to raise and discuss these ideas relatively early, because they require a few key actors (such as technology companies and the AI researchers working for them) to take actions whose benefits will accrue to a much wider population – potentially all of humanity. Thus, we should expect that the default incentives at play will lead to underinvestment in alignment research.[42] The earlier we can understand the risks involved, and the possible ways to avoid them, the easier it will be to build a consensus about the best path forward which is strong enough to overcome whatever self-interested or competitive incentives push in other directions. So despite the inherent difficulty of making arguments about how technological progress will play out, further research into these ideas seems vital for reducing the risk of humanity being left unprepared for the development of AGI.

[42] S Armstrong, N Bostrom, and C Shulman, 'Racing to the Precipice: A Model of Artificial Intelligence Development' (2016) 31 *AI & Society* 201.

3

Artificial Moral Agents

Conceptual Issues and Ethical Controversy

Catrin Misselhorn

I. ARTIFICIAL MORALITY AND MACHINE ETHICS

Artificial Intelligence (AI) has the aim to model or simulate human cognitive capacities. Artificial Morality is a sub-discipline of AI that explores whether and how artificial systems can be furnished with moral capacities.[1] Its goal is to develop artificial moral agents which can take moral decisions and act on them. Artificial moral agents in this sense can be physically embodied robots as well as software agents or 'bots'.

Machine ethics is the ethical discipline that scrutinizes the theoretical and ethical issues that Artificial Morality raises.[2] It involves a meta-ethical and a normative dimension.[3] Meta-ethical issues concern conceptual, ontological, and epistemic aspects of Artificial Morality like what moral agency amounts to, whether artificial systems can be moral agents and, if so, what kind of entities artificial moral agents are, and in which respects human and artificial moral agency diverge.

Normative issues in machine ethics can have a narrower or wider scope. In the narrow sense, machine ethics is about the moral standards that should be implemented in artificial moral agents, for instance: should they follow utilitarian or deontological principles? Does a virtue ethical approach make sense? Can we rely on moral theories that are designed for human social life, at all, or do we need new ethical approaches for artificial moral agents? Should artificial moral agents rely on moral principles at all or should they reason case-based?

In the wider sense, machine ethics comprises the deliberation about the moral implications of Artificial Morality on the individual and societal level. Is Artificial Morality a morally good thing at all? Are there fields of application in which artificial moral agents should not be deployed, if they should be used at all? Are there moral decisions that should not be delegated to machines? What is the moral and legal status of artificial moral agents? Will artificial moral agents change human social life and morality if they become more pervasive?

This article will provide an overview of the most central debates about artificial moral agents. The following section will discuss some examples for artificial moral agents which show that the topic is not just a problem of science fiction and that it makes sense to speak of artificial agents. Afterwards, a taxonomy of different types of moral agents will be introduced that helps to understand the aspirations of Artificial Morality. With this taxonomy in mind, the conditions

[1] C Misselhorn, 'Artificial Morality: Concepts, Issues and Challenges' (2018) 55 *Society* 161 (hereafter Misselhorn, 'Artificial Morality').

[2] SL Anderson, 'Machine Metaethics' in M Anderson and SL Anderson (eds), *Machine Ethics* (2011) 21–27.

[3] C Misselhorn, 'Maschinenethik und Philosophie' in O Bendel (ed), *Handbuch Maschinenethik* (2018) 33–55.

for artificial moral agency in a functional sense will be analyzed. The next section scrutinizes different approaches to implementing moral standards in artificial systems. After these narrow machine ethical considerations, the ongoing controversy regarding the moral desirability of artificial moral agents is going to be addressed. At the end of the article, basic ethical guidelines for the development of artificial moral agents are going to be derived from this controversy.

II. SOME EXAMPLES FOR ARTIFICIAL MORAL AGENTS

The development of increasingly intelligent and autonomous technologies will eventually lead to these systems having to face moral decisions. Already a simple vacuum cleaner like Roomba is, arguably, confronted with morally relevant situations. In contrast to a conventional vacuum cleaner, it is not directly operated by a human being. Hence, it is to a certain degree autonomous. Even such a primitive system faces basic moral challenges, for instance: should it vacuum and hence kill a ladybird that comes in its way or should it pull around it or chase it away? How about a spider? Should it extinguish the spider or save it?

One might wonder whether these are truly moral decisions. Yet, they are based on the consideration that it is wrong to kill or harm animals without a reason. This is a moral matter. Customary Roombas do, of course, not have the capacity to make such a decision. But there are attempts to furnish a Roomba prototype with an ethics module that does take animals' lives into account.[4] As this example shows, artificial moral agents do not have to be very sophisticated and their use is not just a matter of science fiction. However, the more complex the areas of application of autonomous systems get, the more intricate are the moral decisions that they would have to make.

Eldercare is one growing sector of application for artificial moral agents. The hope is to meet demographic change with the help of autonomous artificial systems with moral capacities which can be used in care. Situations that require moral decisions in this context are, for instance: how often and how obtrusively should a care system remind somebody of eating, drinking, or taking a medicine? Should it inform the relatives or a medical service if somebody has not been moving for a while and how long would it be appropriate to wait? Should the system monitor the user at all times and how should it proceed with the collected data? All these situations involve a conflict between different moral values. The moral values at stake are, for instance, autonomy, privacy, physical health, and the concerns of the relatives.

Autonomous driving is the application field of artificial moral agents that probably receives the most public attention. Autonomous vehicles are a particularly delicate example because they do not just face moral decisions but moral dilemmas. A dilemma is a situation in which an agent has two (or more) options which are not morally flawless. A well-known example is the so-called trolley problem which goes back to the British philosopher Philippa Foot.[5] It is a thought experiment which is supposed to test our moral intuitions on the question whether it is morally permissible or even required to sacrifice one person's life in order to save the lives of several persons.

Autonomous vehicles may face structurally similar situations in which it is inevitable to harm or even kill one or more persons in order to save others. Suppose a self-driving car cannot stop and it has only the choice to run into one of two groups of people: on the one hand, two elderly

[4] O Bendel, 'Ladybird: The Animal-Friendly Robot Vacuum Cleaner' (2017) *The AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01* 2-6.

[5] P Foot, *The Problem of Abortion and the Doctrine of Double Effect. Virtues and Vices* (1978) 19–32.

men, two elderly women and a dog; on the other hand, a young woman with a little boy and a little girl. If it hits the first group the two women will be killed, the two men and the dog are going to be severely injured. If it runs into the second group one of the children will get killed and the woman and the other child will be severely injured.

More details can be added to the situation at will. Suppose the group of the elderly people with the dog behaves in accord with the traffic laws, whereas the woman and the children cross the street against the red light. Is this morally relevant? Would it change the situation if one of the elderly men is substituted by a young medical doctor who might save many people's lives? What happens if the self-driving car can only save the life of other traffic participants by sacrificing its passengers?[6] If there is no morally acceptable solution to these dilemmas, this might become a serious impediment for fully autonomous driving.

As these examples show, a rather simple artificial system like a vacuuming robot might already face moral decisions. The more intelligent and autonomous these technologies get, the more intricate the moral problems they confront will become; and there are some doubts as to whether artificial systems can make moral decisions which require such a high degree of sophistication, at all, and whether they should do so.

One might object that it is not the vacuuming robot, the care system, or the autonomous vehicle that makes a moral decision in these cases but rather the designers of these devices. Yet, progress in artificial intelligence renders this assumption questionable. AlphaGo is an artificial system developed by Google DeepMind to play the board game Go. It was the first computer program to beat some of the world's best professional Go players on a full-sized board. Go is considered an extremely demanding cognitive game which is more difficult for artificial systems to win than other games such as chess. Whereas AlphaGo was trained with data from human games; the follow-up version AlphaGoZero was completely self-taught. It came equipped with the rules of the game and perfected its capacities by playing against itself without relying on human games as input. The next generation was MuZero which is even capable of learning different board games without being taught the rules.

The idea that the designers can determine every possible outcome already proves inadequate in the case of less complex chess programs. The program is a far better chess player than its designers who could certainly not compete with the world champions in the game. This holds true all the more for Go. Even if the programmers provide the system with the algorithms on which it operates, they cannot anticipate every single move. Rather, the system is equipped with a set of decision-making procedures that enable it to make effective decisions by itself. Due to the lack of predictability and control by human agents, it makes sense to use the term 'artificial agent' for this kind of system.

III. CLASSIFICATION OF ARTIFICIAL MORAL AGENTS

Even if one agrees that there can be artificial moral agents, it is clear that even the most complex artificial systems differ from human beings in important respects that are central to our

[6] One can find these and some more morally intricate scenarios for self-driving vehicles at http://moralmachine.mit.edu/. The website was created by the MIT with the aim of providing a platform for '1) building a crowd-sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas, and 2) crowd-sourcing assembly and discussion of potential scenarios of moral consequence.' The results were published in different papers that are available at the website.

understanding of moral agency. It is, therefore, common in machine ethics to distinguish between different types of moral agents depending on how highly developed their moral capacities are.[7]

One influential classification of moral agents goes back to *James H. Moor*.[8] He suggested a hierarchical distinction between four types of ethical agents.[9] It does not just apply to artificial systems but helps to understand which capacities an artificial system must have in order to count as a moral agent, although it might lack certain capacities which are essential to human moral agency.

The most primitive form describes agents who generate moral consequences without the consequences being intended as such. *Moor* calls them ethical impact agents. In this sense, every technical device is a moral agent that has good or bad effects on human beings. An example for an ethical impact agent is a digital watch that reminds its owners to keep their appointments on time. However, the moral quality of the effects of these devices lies solely in the use that is made of them. It is, therefore, doubtful whether these should really be called agents. In the case of these devices, the term 'operational morality,' which goes back to *Wendell Wallach* and *Colin Allen*, seems to be more adequate since it does not involve agency.[10]

The next level is taken by implicit ethical agents, whose construction reflects certain moral values, for example security considerations. For *Moor*, this includes warning systems in aircrafts that trigger an alarm if an aircraft comes too close to the ground or if a collision with another aircraft is imminent. Another example are ATMs: these machines do not just have to always emit the right amount of money; they often also check whether money can be withdrawn from the account on that day at all. *Moor* even goes so far as to ascribe virtues to these systems that are not acquired through socialization, but rather directly grounded in the hardware. Conversely, there are also implicit immoral agents with built-in vices, for example a slot machine that is designed in such a way that people invest as much time and money as possible in it. Yet, as in the case of ethical impact agents these devices do not really possess agency since their moral qualities are entirely due to their designers.

The third level is formed by explicit ethical agents. In contrast to the two previous types of agents, these systems can explicitly recognize and process morally relevant information and come to moral decisions. One can compare them to a chess program: such a program recognizes the information relevant to chess, processes it, and makes decisions, with the goal being to win the game. It represents the current position of the pieces on the chessboard and can discern which moves are allowed. On this basis, it calculates which move is most promising under the given circumstances.

For *Moor*, explicit moral agents act not only in accordance with moral guidelines, but also on the basis of moral considerations. This is reminiscent of *Immanuel Kant's* distinction between action in conformity with duty and action from duty.[11] Of course, artificial agents cannot strictly be moral agents in the *Kantian* sense because they do not have a will and they do not have inclinations that can conflict with the moral law. Explicit moral agents are situated somewhere

[7] For an overview, see JA Cervantes and others, 'Artificial Moral Agents: A Survey of the Current Status' (2020) 26 *Science and Engineering Ethics* 501–532.

[8] JH Moor, 'The Nature, Importance, and Difficulty of Machine Ethics' (2006) 21 *IEEE Intelligent Systems* 18–21.

[9] Moor uses the terms 'ethical' and 'moral' synonymously. I prefer to distinguish between these two terms. According to my understanding, morality is the object of ethics. It refers to a specific set of actions, norms, sentiments, attitudes, decisions, and the like. Ethics is the philosophical discipline that scrutinizes morality.

[10] This will be spelled out in the next section. W Wallach and C Allen, *Moral Machines: Teaching Robots Right from Wrong* (2009) 26 (hereafter Wallach and Allen, *Moral Machines*).

[11] M Gregor (ed), *Immaul Kant: Groundwork of the Metaphysics of Morals* (1996) 4:397f.

in between moral subjects in the *Kantian* sense, who act from duty, and *Kant's* example of the prudent merchant whose self-interest only accidentally coincides with moral duty. What *Moor* wants to express is that an explicit moral agent can discern and process morally relevant aspects as such and react in ways that fit various kinds of situations.

Yet, *Moor* would agree with *Kant* that explicit moral agents still fall short of the standards of full moral agency. *Moor's* highest category consists of full ethical agents who have additional capacities such as consciousness, intentionality, and free will, which so far only human beings possess. It remains an open question whether machines can ever achieve these properties. Therefore, *Moor* recommends viewing explicit moral agents as the appropriate target of Artificial Morality. They are of interest from a philosophical and a practical point of view, without seeming to be unrealistic with regard to the technological state of the art.

Moor's notion of an explicit ethical agent can be explicated with the help of the concept of functional morality introduced by *Wallach* and *Allen*.[12] They discriminate different levels of morality along two gradual dimensions: autonomy and ethical sensitivity. According to them, Moor's categories can be situated within their framework.

A simple tool like a hammer possesses neither autonomy nor ethical sensitivity. It can be used to bang a nail or to batter somebody's skull. The possibility of a morally beneficial or harmful deployment would, in Moor's terminology, arguably justify calling it an ethical impact agent, but the artefact as such does not have any moral properties or capacity to act. A child safety lock in contrast does involve a certain ethical sensitivity despite lacking autonomy. It would fall into *Moor's* category of an implicit ethical agent. Because its ethical sensitivity is entirely owed to the design of the object *Wallach* and *Allen* avoid the term of agency and speak of operational morality.

Generally, autonomy and ethical sensitivity are independent of each other.[13] There are, on the one hand, systems which possess a high degree of autonomy, but no (or not much) ethical sensitivity, for example an autopilot. On the other hand, there are systems with a high degree of ethical sensitivity, but no (or a very low degree of) autonomy, for example the platform 'MedEthEx' which is a computer-based learning program in medical ethics.[14] 'MedEthEx' as well as the autopilot belong to the category of functional morality for *Wallach* and *Allen*. Functional morality requires that a machine has 'the capacity for assessing and responding to moral challenges'.[15] This does not necessarily seem to involve agency. If this is the case, there is a level of functional morality below the level of moral agency.[16] Therefore, it has to be specified in more detail which conditions a functional artificial moral agent has to meet.

IV. ARTIFICIAL SYSTEMS AS FUNCTIONAL MORAL AGENTS

There seems to be an intuitive distinction between the things that merely happen to somebody or something and the things that an agent genuinely does.[17] The philosophical question is how to distinguish an action from a mere happening or occurrence and which capacities an object must

[12] Wallach and Allen, *Moral Machines* (n 10) 26.
[13] Ibid, (n 10) 32.
[14] M Anderson, SL Anderson, and C Armen, 'MedEthEx: A Prototype Medical Ethics Advisor' (2006) Proceedings of the Eighteenth Innovative Applications of Artificial Intelligence Conference.
[15] Wallach and Allen, *Moral Machines* (n 10) 9.
[16] Ibid. (n 10) 27.
[17] E Himma, 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?' (2009) 11 *Ethics and Information Technology* 19–29 (hereafter Himma, 'Artificial Agency'); G Wilson and S Shpall, 'Action' in EN Zalta (ed), *Stanford Encyclopedia of Philosophy* (2012).

have in order to qualify as an agent. The range of behaviors that count as actions is fairly broad. It starts from low-level cases of purposeful animal behavior like a spider walking across the table and extends to high-level human cases involving intentionality, self-consciousness, and free will.[18]

A minimal condition for agency is interactivity, i.e. 'that the agent and its environment [can] act upon each other.'[19] Yet, interactivity is not sufficient for agency. The interactions of an agent involve a certain amount of autonomy and intelligence which can vary in degree and type.

The view is expressed, for instance, by the following definition of an artificial agent:

> The term agent is used to represent two orthogonal entities. The first is the agent's ability for autonomous execution. The second is the agent's ability to perform domain-oriented reasoning.[20]

The term 'autonomous execution' means that, although the system is programmed, it acts in a specific situation without being operated or directly controlled by a human being. A higher degree of autonomy arises if a system's behavior becomes increasingly flexible and adaptive, in other words, if it is capable of changing its mode of operation or learning.[21]

Different natural and artificial agents can be situated at different levels of agency depending on their degree and type of autonomy and intelligence. They can, for instance, be classified as goal-directed agents, intentional agents, agents with higher order intentionality, or persons.[22] Distinctive of moral agency is a special kind of domain-oriented reasoning. Explicit ethical agents in *Moor's* sense of the term would have to be able to act from moral reasons.

According to the philosophical standard theory which goes back to *David Hume*, a reason for an action consists in a combination of two mental attitudes: a belief and a pro-attitude. A belief consists in holding something true; a pro-attitude indicates that something ought to be brought about that is not yet the case. Desires are typical pro-attitudes. For this reason, the approach is also often called Belief-Desire-Theory. Take an example: The reason for my action of going to the library may be my desire to read *Leo Tolstoy's* novel 'Anna Karenina', together with the belief that I will find the book in the library. Some versions of the standard theory assume that action explanation also has to refer to an intention that determines which desire will become effective and that includes some plan of action.[23] This accommodates the fact that we have a large number of noncommittal desires that do not lead to actions.[24]

A moral action can thus be traced back to a moral reason, in other words to some combination of moral pro-attitude and corresponding belief. A moral reason may comprise, for instance, the utilitarian value judgment that it is good to maximize pleasure (pro-attitude) and the belief that making a donation to a charitable organization will result in the overall best balance of pleasure versus pain.[25]

[18] H Frankfurt, 'The Problem of Action'(1978) 15 *American Philosophical Quarterly*, 157–162.

[19] L Floridi and JW Sanders, 'On the Morality of Artificial Agents'(2004). 14 *Minds and Machines*, 349, 357 (hereafter Floridi and Sanders, 'On the Morality of Artificial Agents').

[20] The MuBot Agent, cited by S Franklin and A Graesser 'Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents' in JP Mueller, MJ Wooldridge and NR Jennings (eds) *Intelligent Agents III Agent Theories, Architectures, and Languages* (1997) 22.

[21] Floridi and Sanders, 'On the Morality of Artificial Agents' (n 19), regard adaptivity as a separate condition of agency in addition to interactivity and basic autonomy. I prefer to describe it as a higher degree of autonomy. But this might just be a terminological difference.

[22] C Misselhorn 'Collective Agency and Cooperation in Natural and Artificial Systems' in C Misselhorn (ed), *Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation* (2015) 3–25.

[23] ME Bratman, *Intention, Plans, and Practical Reason* (1987).

[24] In the following, this complication is set aside for the sake of simplicity.

[25] It is assumed that we have an intuitive grasp of what moral judgements are. More explicit criteria are given in C Misselhorn, *Grundfragen der Maschinenethik* (4th ed. 2020) (hereafter Misselhorn, *Grundfragen*).

It is a matter of controversy whether artificial systems can possess mental states such as beliefs and desires. Some authors argue that this is not the case because artificial systems do not have intentionality. Intentionality in this sense refers to the fact that mental states like beliefs and desires are about or represent objects, properties, or states of affairs. Most famously *Donald Davidson* assumed that intentionality presupposes complex linguistic abilities which, only humans have.[26] Others concede that animals might also possess intentional states like beliefs and desires, although they do not meet *Davidson's* strong requirements for rationality.[27] This seems to bring intentional agency within the reach of artificial systems as well.[28]

Which stand one takes on this issue depends on the conditions that have to be fulfilled in order to attribute beliefs and desires to an artificial system. According to an instrumentalist view which is often ascribed to *Daniel Dennett*, attributing intentional states is just an explanatory strategy. He argues that states like beliefs and desires are attributed to an agent if this assumption helps us to better understand its behavior, independently of whether there are any corresponding inner states. *Dennett* calls this the intentional stance and the systems that can thus be explained intentional systems. What matters is that we can explain and predict a system's behavior fruitfully by ascribing intentional states to it:

> The success of the stance is of course a matter settled pragmatically, without reference to whether the object really has beliefs, intentions, and so forth; so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably are intentional systems, for they are systems whose behavior can be predicted, and most efficiently predicted, by adopting the intentional stance toward them.[29]

Rational agency is thus a matter of interpretation and does not require that an entity actually possesses internal states, such as beliefs and desires. This condition can be satisfied by artificial systems. For example, if we can understand a chess computer by assuming that it wants to win the game and thinks that a certain move is appropriate to do so, then we can attribute the appropriate reason for action to the computer. Although the behavior of the computer could, in principle, be explained in purely physical terms, the intentional stance is particularly helpful with regard to complex systems.

In contrast, non-instrumental views are not satisfied with reducing intentionality to an attributional practice. Rather, an entity must have certain internal states that are functionally equivalent to beliefs and pro-attitudes.[30] If an artificial system possesses states which have an analogous function for the system as the corresponding mental states have in humans, the system may be called functionally equivalent to a human agent in this respect.

Since there are different ways of specifying the relevant functional relations, functional equivalence has to be seen relative to the type of functionalism one assumes. The most straightforward view with regard to Artificial Morality is machine functionalism which equates the mind directly with a Turing machine whose states can be specified by a machine table. Such a machine table consists of conditionals of the form: 'if the machine is in state S_i and receives input I_j it emits output O_k and goes into state S_l.'[31]

[26] D Davidson, 'Rational Animals' (1982) 36 *Dialectica* 317–327.
[27] F Dretske, *Explaining Behavior: Reasons in a World of Causes* (4th printing 1995).
[28] Dretske remained, however, skeptical with regard to the possibility of obtaining genuine AI as long as artificial systems lack the right kind of history; see F Dretske, 'Can Intelligence Be Artificial?' (1993) 71 *Philosophical Studies* 201–216.
[29] D Dennett, 'Mechanism and Responsibility' in T Honderich (ed), *Essays on Freedom of Action* (1973) 164–165.
[30] This view can also be used to characterize the intentional states of group agents, see C List and P Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Agents* (2011).
[31] N Block, 'Troubles with Functionalism' (1978) 9 *Minnesota Studies in the Philosophy of Science* 261, 266.

Analytic functionalism specifies the relevant functional relations by the causal role of mental terms in folk psychology and rests on the analysis of the meanings of mental terms in ordinary language. Psycho-functionalism, in contrast, defines mental states by their functional role in scientific psychology. This leads to different ways of specifying the relevant inputs, outputs, and internal relations. Analytic functionalism relies on externally observable inputs and outputs, in other words, objects which are located in the vicinity of an organism and bodily movements, as well as common sense views about the causal relations between mental states. Psycho-functionalism can, in contrast, describe functional relations at a neuronal level.

The different types of functionalism also differ with respect to the granularity of their descriptions of the structure of mental states. Simple machine functionalism, for instance, takes mental states like beliefs or desires as unstructured entities. The representational theory of the mind, in contrast, regards mental states as representations with an internal structure that explains the systematic relations between them and the possibility to form indefinitely many new thoughts. The thought 'John loves Mary' has, for instance, the components 'John', 'loves' and 'Mary' as its constituents that can be combined to form other thoughts like 'Mary loves John'.

The most famous proponent who combines a representational view with a computational theory of the mind is *Jerry Fodor*. He regards mental processes as Turing-style computations that operate over structured symbols which are similar to expressions in natural language and form a 'language of thought'.[32] According to *Fodor* and a number of other cognitive scientists, Turing-style computation over mental symbols is 'the only game in town', in other words the only theory that can provide the foundations for a scientific explanation of the mind in cognitive science.[33]

Although the computational model of the mind became enormously influential in the philosophy of mind and cognitive science, it has also been severely criticized. One of the most famous objections against it was developed by *John Searle* with the help of the thought experiment of the Chinese Room.[34] It is supposed to show that Turing-style computation is not sufficient for thought. *Searle* imagines himself in a room manually executing a computer program. Chinese symbols, that people from outside the room slide under the door, represent the input. *Searle* then produces Chinese symbols as an output on the basis of a manual of rules that links input and output without specifying the meaning of the signs. Hence, he produces the appearance of understanding Chinese by following a symbol processing program but does not actually have any language proficiency in Chinese. Because he does not know Chinese, these symbols are only meaningless squiggles to him. Yet, his responses make perfect sense to the Chinese people outside the room. The thought experiment is supposed to trigger the intuition that the system clearly does not understand Chinese, although its behavior is from the outside indistinguishable from a Chinese native speaker. One might also understand the argument as making the point that syntax is not sufficient for semantics, and that computers will never have genuine understanding viz. intentionality because they can only operate syntactically.

If *Searle* is right, machines cannot really possess mental states. They might, however, exhibit states that are functionally equivalent to mental states although they are not associated with phenomenal consciousness and have only derived intentionality mediated by their programmers and users. One might call such states quasi-beliefs, quasi-desires, etc.[35] This way of speaking borrows from the terminology of *Kendall Walton*, who calls emotional reactions to fiction

[32] JA Fodor, *The Language of Thought* (1975).
[33] For a critical assessment of this claim see E Thompson, *Mind in Life* (2007).
[34] JR Searle, 'Minds, Brains, and Programs' (1980) 3 *The Behavioral and Brain Sciences* 417.
[35] Misselhorn, *Grundfragen* (n 25) 86.

(for example, our pity for the protagonist of the novel 'Anna Karenina') quasi-emotions.[36] This is because they do resemble real emotions in terms of their phenomenal quality and the bodily changes involved: we weep for Anna Karenina and feel sadness in the face of her fate. Unlike genuine emotions, quasi-emotions do not involve the belief that the object that triggers the emotion exists.

With artificial moral agents, it is the other way around. They possess only quasi-intentional states that are, unlike their genuine counterparts, not associated with phenomenal consciousness and have only derived intentionality to speak with *Searle* again. For an explicit moral agent in the sense specified above with regard to *Moore's* classification of artificial moral agents, it seems to be sufficient to have such quasi-intentional states. Given the gradual view of moral agency that was introduced in this section, these agents may be functional moral agents although they are not full moral agents on a par with human beings. Arguments to the effect that artificial systems cannot be moral agents at all because they lack consciousness or free will are, hence, falling short.[37]

Functional moral agents are, however, limited in two ways. First, the functional relations just refer to the cognitive aspect of morality. The emotional dimension could be considered only insofar as emotions can be functionally modelled independently of their phenomenal quality. Secondly, functional equivalence is relative to the type of functionalism embraced and functional moral agents possess (so far) at most a subset of the functional relations that characterize full human moral agents. This holds all the more since artificial system's moral reasoning is to date highly domain specific.

It is also important to stress that the gradual view of agency does not imply that functional moral agents are morally responsible for their doings. From a philosophical point of view, the attribution of moral responsibility to an agent requires free will and intentionality.[38] These conditions are not met in the case of functional moral agents. Hence, they do not bear moral responsibility for their doings.

The most fruitful view for the design of artificial moral agents thus lies somewhere in between *Dennett's* instrumentalist conception, which largely abstracts from the agent's internal states, and computational functionalism as a reductive theory of the mind.[39] *Dennett* makes it too easy for machines to be moral agents. His position cannot provide much inspiration for the development of artificial moral agents because he sees the machine merely as a black box; *Fodor's* psycho-functionalism, on the other hand, makes it extremely difficult.

V. APPROACHES TO MORAL IMPLEMENTATION: TOP-DOWN, BOTTOM-UP, AND HYBRID

Moral implementation is the core of Artificial Morality.[40] It concerns the question of how to proceed when designing an artificial moral agent. One standardly distinguishes between

[36] K Walton, 'Fearing Fictions' (1978) 75 *The Journal of Philosophy* 5.

[37] Himma, 'Artificial Agency' (n 17).

[38] F Rudy-Hiller, 'The Epistemic Condition for Moral Responsibility' (2018) in EN Zalta (ed), *Stanford Encyclopedia of Philosophy*.

[39] C Allen, *Intentionality: Natural and Artificial* (2001), even suggests to regard the concept of intentionality as relative to certain explanatory purposes.

[40] For the following, see C Misselhorn, 'Artificial Systems with Moral Capacities? A Research Design and its Implementation in a Geriatric Care System' (2020) 278 *Artificial Intelligence* https://philpapers.org/rec/MISASW https://dl.acm.org/doi/abs/10.1016/j.artint.2019.103179 (hereafter Misselhorn, 'Artificial Systems with Moral Capacity'). This article also specifies a methodological framework for implementing moral capacities in artificial systems.

top-down, bottom-up, and hybrid approaches.[41] All three methods bring together a certain ethical view with a certain approach to software design.

Top-down approaches combine an ethical view that regards moral capacities as an application of moral principles to particular cases with a top-down approach to software design. The basic idea is to formulate moral principles like *Kant's* categorical imperative, the utilitarian principle of maximizing utility, or *Isaac Asimov's* three laws of robotics as rules in a software which is then supposed to derive what has to be morally done in a specific situation. One of the challenges that such a software is facing is how to get from abstract moral principles to particular cases. Particularly with respect to utilitarian systems, the question arises as to how much information they should take into account as 'the consequences of an action are essentially unbounded in space and time'.[42] Deontological approaches might, in contrast, require types of logical inference which lead to problems with decidability.[43]

A more fundamental objection against top-down approaches regarding Artificial Morality is the so-called frame problem. Originally, the frame problem referred to a technical problem in logic-based AI. Intuitively speaking, the issue is sorting out relevant from irrelevant information. In its technical form, the problem is that specifying the conditions which are affected by a system's actions does not, in classical logic, license an inference to the conclusion that all other conditions remain fixed. Although the technical problem is largely considered as solved (even within strictly logic-based accounts), there remains a wider, philosophical version of the problem first stated by *John McCar*thy and *Patrick Hayes* which is not yet close to a solution.[44]

The challenge is that potentially every new piece of information may have an impact on the whole cognitive system of an agent. This observation has been used as evidence against a computational approach to the mind because it seems to imply that central cognitive processes cannot be modelled by strictly general rules. A corresponding line of argument can also be turned against top-down approaches regarding Artificial Morality. As *Terry Horgan* and *Mark Timmons* point out, moral normativity is not fully systematizable by exceptionless general principles because of the frame problem.[45] Full systematizability is, however, not required for Artificial Morality, and *Horgan* and *Timmons* admit that a partial systematization of moral normativity via moral principles remains possible. The frame problem is, hence, not a knock-down argument against the possibility of top-down approaches to moral implementation although it remains a challenge for AI in general.

The alternative to top-down are bottom-up approaches which do not understand morality as rule-based. This view is closely related to moral particularism, a meta-ethical position that rejects the claim that there are strict moral principles and that moral capacities consist in the application of moral principles to particular cases.[46] Moral particularists use to think of moral capacities in terms of practical wisdom or in analogy to perception as attending to the morally relevant features (or values) that a situation instantiates. Moral perception views emphasize the individual

[41] Wallach and Allen, *Moral Machines* (n 10).
[42] Wallach and Allen, *Moral Machines* (n 10) 86.
[43] TM Powers, 'Prospects for a Kantian Machine' in M Anderson and SL Anderson (eds), *Machine Ethics* (2011) 464.
[44] J McCarthy and PJ Hayes, 'Some Philosophical Problems from the Standpoint of Artificial Intelligence' in B Meltzer and D Michie (eds), *Machine Intelligence* (1969) 463; M Shanahan, 'The Frame Problem' in EN Zalta (ed), *The Stanford Encyclopedia of Philosophy* (2009).
[45] T Horgan and M Timmons 'What Does the Frame Problem Tell Us About Moral Normativity?' (2009) 12 *Ethical Theory and Moral Practice* 25.
[46] J Dancy, *Ethics Without Principles* (2004).

sensibility to the moral aspects of a situation.[47] The concept of practical wisdom goes back to *Aristotle* who underlined the influence of contextual aspects which are induced by way of socialization or training. In order to bring these capacities about in artificial systems, bottom-up approaches in software design which start from finding patterns in various kinds of data have to be adapted to the constraints of moral learning. This can be done either with the help of an evolutionary approach or by mimicking human socialization.[48]

Bottom-up approaches might thus teach us something about the phylo- and ontogenetical evolution of morality.[49] But, they are of limited suitability for implementing moral capacities in artificial systems because they pose problems of operationalization, safety, and acceptance. It is difficult to evaluate when precisely a system possesses the capacity for moral learning and how it will, in effect, evolve. Because the behavior of such a system is hard to predict and explain, bottom-up approaches are hardly suitable for practical purposes; they might put potential users at risk. Moreover, it is difficult to reconstruct how a system arrived at a moral decision. Yet, it is important that autonomous artificial systems do not just behave morally, as a matter of fact, but that the moral basis of their decisions is transparent. Bottom-up approaches should, as a consequence, be restricted to narrowly confined and strictly controlled laboratory conditions.

Top-down and bottom-up are the most common ways to think about the implementation of moral capacities in artificial systems. It is, however, also possible to combine the virtues of both types of approaches. The resulting strategy is called a hybrid approach. Hybrid approaches operate on the basis of a predefined framework of moral values which is then adapted to specific moral contexts by learning processes.[50] Which values are given depends on the area of deployment of the system and its moral characteristics. Although hybrid approaches are promising, they are still in the early stages of development. So, which approach to moral implementation should one choose? It does not make much sense to answer this question in the abstract. It depends on the purpose and context of use for which a system is designed. An autonomous vehicle will demand a different approach to moral implementation than a domestic care robot.

VI. ETHICAL CONTROVERSY ABOUT ARTIFICIAL MORAL AGENTS

Machine ethics, however, does not just deal with issues about moral agency and moral implementation. It also discusses the question of whether artificial moral agents should be approved from a moral point of view. This became a major topic in the last years because Artificial Morality is part of technological innovations that are disruptive and can change individual lives and society profoundly. Not least, a lot of effort and money is spent on research on artificial moral agents in different domains, which also receives a lot of public and media attention. A number of big companies and important economic players strongly push Artificial Morality in areas like autonomous driving, and politics removes, under the perceived economic pressure, more and more legal barriers that might so far prevent the commercial launch of these technologies.

[47] M Nussbaum, 'Finely Aware and Richly Responsible: Moral Attention and the Moral Task of Literature'(1985) 82 *Journal of Philosophy* 516.

[48] For the first approach, see T Froese and E Di Paolo, 'Modelling Social Interaction as Perceptual Crossing: An Investigation into the Dynamics of the Interaction Process' (2010) 22 *Connection Science* 43; for the second, see C Breazeal and B Scassellati, 'Robots That Imitate Humans' (2002) 6 *Trends in Cognitive Sciences* 481; T Fong, N Illah, and K Dautenhahn, 'A Survey of Socially Interactive Robots: Concepts, Design, and Applications' (2002) CMU-RI-TR Technical Report 2.

[49] R Axelrod, *The Evolution of Cooperation* (1984).

[50] For a hybrid approach to a software module for an elder care system, see Misselhorn, 'Artificial Systems with Moral Capacity' (n 40).

The ethical evaluation ranges from a complete refusal of artificial moral agents, over balanced assessments stressing that the moral evaluation of Artificial Morality has to take into account the diversity of approaches and application contexts, to arguments for the necessity of artificial moral agents.[51] The following overview tries to take up the most salient issues but it does not intend to be exhaustive. It focusses on questions that arise specifically with respect to artificial moral agents and does not comment on topics like privacy that belong to the more generic discipline of ethics of AI.

1. *Are Artificial Moral Agents Inevitable?*

One important argument in the discussion is that artificial moral agents are inevitable.[52] The development of increasingly complex intelligent and autonomous technologies will eventually lead to these systems having to face morally problematic situations which cannot be fully controlled by human operators. If this is true, the need for artificial moral agents is eventually arising from technological progress. It would, however, be wrong to either accept this development fatalistically or to reject it as such, because inevitability is a conditional matter. If we want to use intelligent and autonomous technologies in certain areas of application, then this will inevitably lead to the need for artificial moral agents. Hence, we should deliberate in which areas of application – if any – it is right from a moral point of view to use such agents and in which areas it would be morally wrong.[53]

2. *Are Artificial Moral Agents Reducing Ethics to Safety?*

Another motivation for building artificial moral agents is a concern with safety. The idea is that equipping machines with moral capacities can prevent them from harming human beings. It would, however, be wrong to reduce ethics to safety issues.[54] There are other important moral values that can conflict with safety and that have to be taken into consideration by an artificial moral agent. In the context of elder care, safety would, for instance, consist in avoiding health risks at all costs. Yet, this might conflict with the caretakers autonomy.[55] Although safety is a moral value that has to be taken into consideration in developing artificial moral agents, Artificial Morality cannot be reduced to it.

3. *Can Artificial Moral Agents Increase Trust in AI?*

A third aspect that is invoked in the discussion is that artificial moral agents will increase public trust in artificial intelligence. The hope is that Artificial Morality might in this way help to deal with the fears that many people feel with regard to artificial intelligence and robots and improve

[51] For the first position, see A Van Wysberghe and S Robbins, 'Critiquing the Reasons for Making Artificial Moral Agents' (2019) 25 *Science and Engineering Ethics* 719 (hereafter Van Wysberghe and Robbins, 'Critiquing the Reasons'). For the intermediate view, see Misselhorn, 'Artificial Morality' (n 1) and for the last view, see P Formosa and M Ryan, 'Making Moral Machines: Why We Need Artificial Moral Agents' (2020) *AI & Society* https://link.springer.com/article/10.1007/s00146-020-01089-6 (hereafter Formosa and Ryan, 'Making Moral Machines').

[52] This claim is defended by, among others, C Allen and W Wallach, 'Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?' in P Lin, K Abney, and GA Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (2011) 55.

[53] Misselhorn, 'Artificial Morality' (n 1).

[54] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51).

[55] Misselhorn, 'Artificial Systems with Moral Capacity' (n 40); Formosa and Ryan, 'Making Moral Machines' (n 51).

the acceptance of these technologies.[56] One must, however, distinguish between trust and reliance.[57] Trust is an emotional attitude that arises in a relationship involving mutual attitudes toward one another which are constitutive.[58] It does, for instance, lead to the feeling of being betrayed and not just disappointed when let down.[59] This presupposes the ascription of moral responsibility that must be denied to functional moral agents as argued above. Hence, we should rather speak of reliance instead of trust in artificial moral agents.

It is, moreover, advisable not to be too credulous with regard to artificial moral agents. The lack of predictability and control invoked before to justify why it is adequate to speak of moral agents is also a good reason for not relying blindly on them. The danger is that the term 'Artificial Morality' is suggestively used to increase unjustified acceptance although we should, from a moral point of view, rather keep a critical eye on artificial moral agents.

Even if artificial moral agents do not fulfill the conditions for trustworthiness, trust may play a role with respect to the design and development of artificial moral agents. Suggestions to ensure trust in these cases include a code of conduct for the designers of these devices, transparency with regard to moral implementation, and design of artificial moral agents, as well as standards and certifications for the development process comparable to FairTrade, ISO, or GMOs.[60] Particularly in areas of application that concern not just the users of artificial moral agents but affect the population more broadly or have a large impact on the public infrastructure, like autonomous driving, it is a political task to establish democratically legitimized laws for the design and development process of artificial moral agents or even to constrain their development if necessary.

4. Do Artificial Moral Agents Prevent Immoral Use by Design?

Another argument in favor of artificial moral agents is that they prevent being used immorally by design. Major objections against this argument are that this massively interferes with the autonomy of human beings and can lead to unfair results. Amazon is, for instance, about to install a system called Driveri in their delivery vehicles in the United States. This is an automated monitoring system that consists of high-tech cameras combined with a software which is used to observe and analyze the drivers' behavior when operating the car. It gives real-time feedback in certain cases, for instance, when the driver is going too fast, seems to be distracted, or does not wear a seatbelt. When it comes to the conclusion that something went badly wrong, it will give the information to actual humans at the company.[61] The data are also used to evaluate the drivers and might lead to them being fired – by a machine. Amazon promotes the system as improving safety. But it is clear that it cannot take the subtleties and complexities of human life into account. Sometimes there are good reasons to deviate from the rules or there are special circumstances that the drivers could not influence. This may lead to unfair decisions and hardships that can destroy lives.[62]

[56] M Anderson, SL Anderson, 'Machine Ethics: Creating an Ethical Intelligent Agent' 28 *AI Magazine* 15.
[57] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51).
[58] C McLeod, 'Trust' in EN Zalta (ed), *Stanford Encyclopedia of Philosophy*.
[59] A Baier, 'Trust and Antitrust'(1986) 96 *Ethics* 231; J Simon, 'The Entanglement of Trust and Knowledge on the Web' (2010) 12 *Ethics and Information Technology* 343.
[60] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51) 728.
[61] J Stanley, 'Amazon Drivers Placed Under Robot Surveillance Microscope' (*ACLU*, 23 March 2021) www.aclu.org/news/privacy-technology/amazon-drivers-placed-under-robot-surveillance-microscope/.
[62] S Soper, 'Fired by Bot at Amazon: "It's You Against the Machine"' (*Bloomberg*, 28 June 2021) www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out.

Consider some other examples: how about a woman who had a couple of drinks with her partner at home and then refuses to have sex with him. Imagine that her partner gets violent and the woman tries to get away by car but the breathalyzer reacts to the alcohol in her breath and does not let her start the car.[63] Is it the right decision from a moral point of view to prevent the woman from driving because she drank alcohol and to expose her to domestic violence? How about elderly persons at home who ask their service robots for another glass of wine or pizza every day? Should the robot deny to get these things if it thinks that they are a health risk for the user as it happens in the Swedish TV-series *Real Humans*? Examples like these show that it is far from clear which uses are strictly immoral and should be precluded by design. One might, of course, try to deal with the problem by giving people always the possibility to override the system's decisions. But that would undermine the whole purpose of preventing immoral uses by design.

5. *Are Artificial Moral Agents Better than Humans?*

A yet stronger claim is that artificial moral agents are even morally better than humans because their behavior is not influenced by irrational impulses, psychopathologies, or emotional distress. They are impartial, not prone to bias, and they are not diverted from the path of virtue by self-interest. Moreover, machines may be superior to humans in their cognitive abilities. They are able to make decisions in fractions of a second, during which a human being cannot come to conscious decisions. This is used as an argument for leaving moral decisions to machines in particularly precarious situations, for example in war.[64]

Apart from the fact that this argument presupposes an idealized view of AI which does, for instance, ignore the problem of algorithmic bias, several objections have been raised against it. Many argue that artificial systems lack important capacities that human moral agents possess. One point is that emotions are vital for moral judgment and reasoning and that artificial moral agents with emotions are 'something not even on the horizon of AI and robotics'.[65]

As explicated above, this point is somewhat simply put. Emotional AI is a strongly emergent research program inspired by the insights of research in psychology and neuroscience on the importance of emotions for intelligent behavior that goes back to the 1980s.[66] As with artificial moral agency, the state of the art consists in trying to model states that are functionally equivalent to emotions at different levels of granularity.[67] There are even attempts to build artificial moral agents with emotional or empathic capacities.[68] The crucial point is not that emotions are out of the reach of AI, it is that moral emotions involve consciousness and that there is serious doubt that consciousness can be computationally modelled. The crucial question is, therefore, whether functional moral agency is achievable without consciousness.

[63] This case is a slight variation of an example from Van Wysberghe and Robbins, 'Critiquing the Reasons' 729 (n 51).
[64] R Arkin, *Governing Lethal Behavior in Autonomous Robots* (2009) (hereafter Arkin, *Governing*).
[65] Van Wysberghe and Robbins, 'Critiquing the Reasons', 730 (n 51).
[66] M Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence and the Future of the Human Mind* (2006); R Picard, *Affective Computing* (1997).
[67] R Reisenzein and others, 'Computational Modeling of Emotion: Toward Improving the Inter- and Intradisciplinary Exchange' (2013) 4 *IEEE Transactions on Affective Computing* 246.
[68] C Balkenius and others, 'Outline of a Sensory-motor Perspective on Intrinsically Moral Agents' (2016) 24 *Adaptive Behavior* 306; C Misselhorn, *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co* (2021).

6. Does Reasonable Pluralism in Ethics Speak against Artificial Moral Agents?

A rather desperate move by the adversaries of Artificial Morality is to mount moral skepticism, subjectivism, or an error-theory of moral judgments against it.[69] It is true, if there is no moral right and wrong that is at least intersubjectively binding or if all moral judgments are false, then the development of artificial moral agents would not make sense from the start. But this strategy overstates the case and cures one evil with a worse one. The fact of reason, as *Kant* called it; our existing moral practice is enough for getting Artificial Morality off the ground if there are no other reasons against it.

Having said this, one still has to take into account the fact that there is no consensus about the correct moral theory, neither in the general public nor among philosophers. *John Rawls* calls this 'the fact of reasonable pluralism' and he thinks that it is due to burdens of judgment that we cannot overcome. Reasonable pluralism is, for him, 'the inevitable long-run result of the powers of human reason at work within the background of enduring free institutions.'[70] The question then is which morality should be implemented in artificial systems.

The answer to this question depends on the context. Service, care, or household robots that only affect one individual could be programmed in a way that responds to the individual moral framework of the user.[71] If a system operates, in contrast, in the public sphere and its decisions inevitably concern the vital interests of other people apart from its user, the system's behavior should be governed by generally binding political and legal regulations. This would hold, for instance, for autonomous driving. Ethical pluralism is no insurmountable obstacle to establishing laws with respect to controversial ethical issues in liberal democracies. Examples that show this are (at least in Germany) abortion or assisted dying. Although not every individual agrees entirely with the legal regulations in these cases, most citizens find them morally acceptable, although they are not immune to change. In 2020, the German Constitutional Court decided in response to a lawsuit of assisted suicide organizations to abrogate the general prohibition of assisted suicide. Of course, things get more complicated as soon as international standards are required.

The issues about abortion or assisted suicide have, moreover, certain characteristics that make it unclear whether they can be applied directly to artificial moral agents. The regulations set limits to the choices of individuals but they do not determine them. Yet, it is questionable whether artificial moral agents could and should have such latitudes or whether this is the privilege of full moral agents. Another important point is the difference between individual choices and laws. An individual might, for instance, decide to save a child instead of an elderly persona in a dilemma situation in autonomous driving but if politics decided to establish algorithms in autonomous vehicles by law that sacrifice elderly people in dilemma situations that seems to be a case of age discrimination.

7. Do Artificial Moral Agents Threaten Our Personal Bonds?

Another worry is that by fixing moral decisions algorithmically, one does not take into account that some situations lie beyond moral justification, as *Bernard Williams* puts it.[72] He argues that

[69] BC Stahl, 'Information, Ethics, and Computers: The Problem of Autonomous Moral Agents' (2004) 14 *Minds and Machines* 67; Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51).
[70] J Rawls, *Political Liberalism* (1993) 4.
[71] Misselhorn, 'Artificial Systems with Moral Capacity' (n 40).
[72] B Williams, 'Persons, Character, and Morality' in W Bernard, *Moral Luck* (1981) 18.

it would be 'one thought too many' if a husband, faced with the possibility of saving either his wife or a stranger, first has to think about whether it is compatible with his moral principles to give preference to his wife.[73] This is not just a matter of acting instinctively rather than on deliberation. It would be just as inappropriate for the husband to consider in advance whether he should save his wife if he were the captain of the ship and two strangers stood against his wife, or if he should save fifty strangers instead of his wife. The crucial point is that conducting these thought experiments would not be appropriate to the special relationship of mutually loving spouses. Such reasoning threatens to alienate us, according to Williams, from our personal bonds with family or friends. The problem is not just that an artificial moral agent could not make such a decision, the problem is that doing so would undermine its impartiality which was one of the main reasons why artificial moral agents might be considered as superior to human moral agents.

8. *Which Impact Does Artificial Morality Have on Ethical Theory?*

Examples like these have an impact on another issue as well. One might argue that Artificial Morality might help us to improve our moral theories. Human ethics is often fragmented and inconsistent. Creating artificial moral agents could contribute to making moral theory more consistent and unified because artificial systems can only operate on such a basis. Yet, the examples discussed raise the question whether it is good that Artificial Morality forces us to take a stance on cases that have so far not been up for decision or to which there are no clear ethical solutions as in the dilemma cases in autonomous driving.The necessity to decide such cases might, on the one hand, contribute to making our moral views more coherent and unified. On the other hand, the fact that Artificial Morality forces us to take a stance in these cases might incur guilt on us by forcing us to deliberately approve that certain people get harmed or even killed in situations like the dilemmas in autonomous driving. There may simply be cases that resist a definite final solution as Artificial Morality requires it. Some have argued that one should use algorithms that select randomly in such situations.[74] Yet, this solution contradicts the fact that in a specific dilemma situation there might well be morally preferable choices in this particular context although they cannot be generalized. What is more, a random-selecting algorithm seems to express an attitude towards human life that does not properly respect its unique value and dignity.[75]

9. *Is It Wrong to Delegate Moral Decision-Making to Artificial Moral Agents?*

There are also worries to the effect that 'outsourcing' moral decisions to machines deprives human beings of a practice that is morally essential for humanity. According to *Aristotle*, acquiring expertise in moral reasoning belongs necessarily to a human being's good life and this requires gaining moral understanding through practice.[76] Delegating moral decision-making to artificial moral agents will reduce the opportunities to exercise this capacity and will

[73] For an argument against utilitarianism in machine ethics that refers to this view, see: C Grau, 'There Is No "I" in "Robot". Robots and Utilitarianism' in M Anderson and SL Anderson, *Machine Ethics* (2011) Fn 2, 451–463.

[74] L Zhao and W Li, '"Choose for No Choose": Random-Selecting Option for the Trolley Problem in Autonomous Driving' in J Zhang and others (eds), *LISS2019* (2019).

[75] Misselhorn, *Grundfragen* (n 25) 195.

[76] Van Wysberghe and Robbins, 'Critiquing the Reasons' 731 (n 51) 731.

lead to a 'de-skilling' of humans with respect to morality.[77] One might rise to this challenge by pointing out that there are still many opportunities for humans to exercise and develop their moral skills.[78]

Yet, there might be a deeper concern that this answer does not address. For *Kant*, being able to act morally is the source of our normative claims towards others. One might interpret this claim as saying that morality is a reciprocal practice between full moral agents that are autonomous in the sense of setting themselves ends and that are able to reason with each other in a distinctly second-personal way.[79] Functional moral agents cannot really take part in such a practice, and one might argue that delegating moral decisions to them violates this moral practice independently of the quantitative question of how often this is done. This is one of the reasons why creating a Kantian artificial moral agent might be contradictory.[80]

10. *Who Is Responsible for the Decisions of Artificial Moral Agents?*

Finally, there is the concern that Artificial Morality might undermine our current practice of responsibility ascription. As was argued above, delegating morally relevant decisions to artificial systems might create so-called responsibility gaps. *Robert Sparrow* who coined this term uses the example of lethal autonomous weapon systems to argue that a responsibility gap arises if such a system violates the ethical or legal norms of warfare and the following conditions are fulfilled: (1) the system was not intentionally programmed to violate the ethical or legal norms of warfare; (2) it was not foreseeable that the use of the lethal autonomous weapon system would lead to this result; and (3) there was no human control over the machine from the start of the operation.[81]

The problem is that if these three conditions are fulfilled, then moral responsibility cannot be attributed to any human when the machine kills humans in conflict with the moral or legal norms of warfare, because no human being had intended it, it was not foreseeable, and nobody had the possibility to prevent the result. Thus, a responsibility gap occurs precisely when the machine itself is not responsible, but its use undermines the terms of attributing responsibility to human beings. For *Sparrow*, this is a reason for rejecting the use of war robots as immoral because, at least when it comes to killing humans, there should always be someone who can be held responsible for the deaths.

VII. CONCLUSION: GUIDELINES FOR MACHINE ETHICS

Which conclusions should we draw from the controversy about artificial moral agents? One suggestion is to place a moratorium on the commercialization of artificial moral agents. The idea is to allow academic research on Artificial Morality while at the same time protecting users, other concerned persons or groups, and society 'from exposure to this technology which poses an existential challenge'.[82] This seems to be at least reasonable as long as we do not have good answers to the challenges and critical questions discussed in the last section.

[77] S Vallor, 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character' (2015) 28 *Philosophy & Technology* 107.
[78] Formosa and Ryan, 'Making Moral Machines' (n 51).
[79] S Darwall, 'Kant on Respect, Dignity, and the Duty of Respect' in M Betzler (ed), *Kant's Ethics of Virtues* (2008).
[80] R Tonkens, 'A Challenge for Machine Ethics' (2009) 19 *Minds and Machines* 421.
[81] R Sparrow, 'Killer Robots' in 24 *Journal of Applied Philosophy* 62.
[82] Van Wysberghe and Robbins, 'Critiquing the Reasons' (n 51) 732.

There are, however, some loopholes that this suggestion does not address. A device like an autonomous car might, as a matter of fact, be designed as an artificial moral agent without being commercialized as such. This is possible because algorithms are often trade secrets. Another challenge is that moral decisions do not always have to be taken explicitly but might be hidden behind other parameters. An algorithm for autonomous driving might, for instance, give priority to the passengers' safety by using certain technical parameters without making it explicit that this puts the risk on more vulnerable traffic participants.

The controversy about artificial moral agents does, however, not necessarily have to be seen as formulating impediments to research and innovation. The arguments might also be regarded as indicators for the directions that research on the design of artificial moral agents and their development should take. The lessons that have to be drawn from the controversy can be condensed in three fundamental guidelines for machine ethics:[83]

(1) Moral machines should promote human autonomy and not interfere with it.
(2) Artificial systems should not make decisions about life and death of humans.
(3) It must be ensured that humans always take responsibility in a substantial sense.

In the light of these three guidelines for machine ethics, there are some areas of application of artificial moral agents that should be viewed critically from a moral point of view. This applies in particular to killer robots, but autonomous driving should also be considered carefully against this background. There is reason to assume that, in order to optimize accident outcomes, it is necessary to specify cost functions that determine who will be injured and killed which bear some similarity to lethal autonomous weapon systems. Legitimate targets would have to be defined for the case of an unavoidable collision, which would then be intentionally injured or even killed.[84] As long as the controversial issues are not resolved, robots should not get a license to kill.[85]

Even if one does not want to hand over decisions about the life and death of human beings to artificial moral agents, there remain areas of application in which they might be usefully employed. One suggestion is a conceptual design of a software module for elder care that can adapt to the user's individual moral value profile through training and permanent interaction and that can, therefore, treat people according to their individual moral value profile.[86] Under the conditions of reasonable pluralism, it can be assumed that users' values with respect to care differ, for example, as to whether more weight should be given to privacy or to avoiding health risks. A care system should be able to weigh these values according to the moral standards of each individual user. In this case, a care system can help people who wish to do so to live longer in their own homes.

Such a system could be compared to an extended moral arm or prosthesis of the users. One could also speak of a moral avatar which might strengthen the care-dependent persons' self-esteem by helping them to live according to their own moral standards. Yet, such a system is only suitable for people who are cognitively capable of making basic decisions about their lives

[83] These guidelines must be understood as addressing specifically the arguments from the controversy. There are other principles of machine ethics, for instance, that the decisions of artificial moral agents should be fair. Such principles arise from general ethical considerations which are not specific to machine ethics.

[84] P Lin, 'Why Ethics Matters for Autonomous Cars' in M Maurer and others (eds), *Autonomous Driving: Technical, Legal and Social Aspects* (2007).

[85] C Misselhorn, 'Lizenz zum Töten für Roboter? "Terror" und das autonome Fahren' in B Schmidt (ed), *Terror: Das Recht braucht eine Bühne. Essays, Hintergründe, Analysen* (2020).

[86] Misselhorn, *Grundfragen* (n 25).

but are so physically limited that they cannot live alone at home without care. It should also be clear that there is no purely technological solution to the shortage of care givers. It is essential to embed these technologies in a social framework. No one should be cared for by robots against their will. The use of care systems must also not lead to loneliness and social isolation among those receiving care.

A very demanding task is to make sure that humans always take responsibility in a substantial sense as the third principle demands. In military contexts, a distinction is made between in-the-loop systems, on-the-loop systems, and out-of-the-loop systems, depending on the role of the human in the control loop.[87] In the case of in-the-loop systems, a human operates the system and makes all relevant decisions, even if it is by remote control. On-the-loop systems are programmed and can operate in real time independent of human intervention. However, the human is still responsible for monitoring the system and can intervene at any time. Out-of-the-loop systems work like on-the-loop systems, but there is no longer any possibility of human control or intervention.

The problem of the responsibility gap appears to be solved if the human remains at least on-the-loop and perhaps even has to agree to take responsibility by pressing a button before putting an artificial system into operation.[88] But how realistic is the assumption that humans are capable of permanent monitoring? Can they maintain attention for that long, and are they ready to decide and intervene in seconds when it matters? If this is not the case, predictability and control would be theoretically possible, but not feasible for humans in reality.

Second, there arise epistemological problems, if the human operators depend on the information provided by the system to analyze the situation. The question is whether the users can even rationally doubt its decisions if they do not have access to independent information. In addition, such a system must go through a series of quality assurance processes during its development. This may also be a reason for users to consider the system's suggestions as superior to their own doubts. Hence, the problem of the responsibility gap also threatens on-the-loop systems and it may even occur when humans remain in-the-loop.[89]

Overall, it seems unfair that the users should assume full responsibility at the push of a button, because at least part of the responsibility, if not the main part, should go to the programmers, whose algorithms are decisive for the system's actions. The users are only responsible in a weaker sense because they did not prevent the system from acting. A suitable approach must take into account the distribution of responsibility which does not make it easier to come to terms with the responsibility gap. One of the greatest challenges of machine ethics is, therefore, to define a concept of meaningful control and to find ways for humans to assume responsibility for the actions of artificial moral agents in a substantial sense.

[87] United States Department of Defense Unmanned Systems Integrated Roadmap FY 2011-2036. Reference Number 11-S-3613. https://irp.fas.org/program/collect/usroadmap2011.pdf

[88] Such a suggestion is, for instance, made by Arkin, *Governing* (n 64).

[89] A Matthias, 'The Responsibility Gap – Ascribing Responsibility for the Actions of Learning Automata' (2004) 6(3) *Ethics and Information Technology* 175–183.

4

Risk Imposition by Artificial Agents

The Moral Proxy Problem

*Johanna Thoma**

I. INTRODUCTION

It seems undeniable that the coming years will see an ever-increasing reliance on artificial agents that are, on the one hand, autonomous in the sense that they process information and make decisions without continuous human input, and, on the other hand, fall short of the kind of agency that would warrant ascribing moral responsibility to the artificial agent itself. What I have in mind here are artificial agents such as self-driving cars, artificial trading agents in financial markets, nursebots, or robot teachers.[1] As these examples illustrate, many such agents make morally significant decisions, including ones that involve risks of severe harm to humans. Where such artificial agents are employed, the ambition is that they can make decisions roughly as good as or better than those that a typical human agent would have made in the context of their employment. Still, the standard by which we judge their choices to be good or bad is still considered human judgement; we would like these artificial agents to serve human ends.[2]

Where artificial agents are not liable to be ascribed true moral agency and responsibility in their own right, we can understand them as acting as proxies for human agents, as making decisions on their behalf. What I will call the 'Moral Proxy Problem' arises because it is often not clear for whom a specific artificial agent is acting as a moral proxy. In particular, we need to decide whether artificial agents should be acting as proxies for what I will call low-level agents – for example individual users of the artificial agents, or the kinds of individual human agents artificial agents are usually replacing – or whether they should be moral proxies for what I will call high-level agents – for example designers, distributors, or regulators, that is, those who can

* I received very helpful feedback on an earlier draft of this paper from Kate Vredenburgh, Silvia Milano, and Johannes Himmelreich. Previous versions of this paper were also presented at the University of York, at the Global Priorities Institute at Oxford, at the Third Workshop on Decision Theory and the Future of Artificial Intelligence at the Australian National University, at the Humanities and Social Sciences Colloquium at the Freiburg Institute for Advanced Studies (FRIAS), and at the Interdisciplinary Research Symposium on Global Perspectives on Responsible AI organised by the FRIAS Saltus Group on Responsible AI. I have benefitted immensely from discussions both at these events, as well as during my time as a visiting FRIAS Fellow in April 2019.

[1] See M Wellman and U Rajan, 'Ethical Issues for Autonomous Trading Agents' (2017) 27 *Minds and Machines* 609; A Sharkey and N Sharkey, 'Granny and the Robots: Ethical Issues in Robot Care for the Elderly' (2012) 14 *Ethics and Information Technology* 27; and A Sharkey, 'Should We Welcome Robot Teachers?' (2016) 18 *Ethics and Information Technology* 283 respectively for critical discussion of these types of agents.

[2] Note that I don't mean to restrict human ends to human interests in a narrow sense here. Insofar as humans can, and often do, have ends that are not speciesist, we can think of artificial agents being deployed to further such ends, for example in wildlife preservation.

potentially control the choice behaviour of many artificial agents at once. I am particularly interested in the Moral Proxy Problem insofar as it matters for decision structuring when making choices about the design of artificial agents. Who we think an artificial agent is a moral proxy for determines from which agential perspective the choice problems artificial agents will be faced with should be framed:[3] should we frame them like the individual choice scenarios previously faced by individual human agents? Or should we, rather, consider the expected aggregate effects of the many choices made by all the artificial agents of a particular type all at once?

Although there are some initial reasons (canvassed in Section 2) to think that the Moral Proxy Problem and its implications for decision structuring have little practical relevance for design choices, in this paper I will argue that in the context of risk the Moral Proxy Problem has special practical relevance. Just like most human decisions are made in the context of risk, so most decisions faced by artificial agents involve risk:[4] self-driving cars can't tell with complete certainty how objects in their vicinity will move, but rather make probabilistic projections; artificial trading agents trade in the context of uncertainty about market movements; and nursebots might, for instance, need to make risky decisions about whether a patient symptom warrants raising an alarm. I will focus on cases in which the artificial agent can assign precise probabilities to the different potential outcomes of its choices (but no outcome is predicted to occur with 100% certainty). The practical design choice I am primarily concerned with here is how artificial agents should be designed to choose in the context of risk thus understood, and in particular whether they should be programmed to be risk neutral or not. It is for this design choice that the Moral Proxy Problem turns out to be highly relevant.

I will proceed by, in Section III, making an observation about the standard approach to artificial agent design that I believe deserves more attention, namely that it implies, in the ideal case, the implementation of risk neutral pursuit of the goals the agent is programmed to pursue. But risk neutrality is not an uncontroversial requirement of instrumentally rational agency. Risk non-neutrality, and in particular risk aversion, is common in choices made by human agents, and in those cases is intuitively neither always irrational, nor immoral. If artificial agents are to be understood as moral proxies for low-level human agents, they should emulate considered human judgements about the kinds of choice situations low-level agents previously found themselves in and that are now faced by artificial agents. Given considered human judgement in such scenarios, often exhibits risk non-neutrality, and in particular risk aversion; artificial agents that are moral proxies for low-level human agents should do so too, or should at least have the capacity to be set to do so by their users.

Things look differently, however, when we think of artificial agents as moral proxies for high-level agents, as I argue in Section IV. If we frame decisions from the high-level agential perspective, the choices of an individual artificial agent should be considered as part of an aggregate of many similar choices. I will argue that once we adopt such a compound framing, the only reasonable approach to risk is that artificial agents should be risk neutral in individual choices, because this has almost certainly better outcomes in the aggregate. Thus, from the high-level agential perspective, the risk neutrality implied by the standard approach appears justified. And so, how we resolve the Moral Proxy Problem is of high practical importance in the context of risk. I will return to the difficulty of addressing the problem in Section V, and also argue there

[3] Here and throughout, I use 'framing' in a non-pejorative sense, as simply referring to the way in which a decision problem is formulated before it its addressed.

[4] Frequently neglecting the context of risk is indeed a serious limitation of many discussions on the ethics of AI. See also S Nyholm and J Smids, 'The Ethics of Accident-Algorithms for Self-Driving Cars; an Applied Trolley Problem?' (2016) 19 *Ethical Theory and Moral Practice* 1275 (hereafter Nyholm and Smids, 'Ethics of Accident-Algorithms').

that the practical relevance of agential framing is problematic for the common view that responsibility for the choices of artificial agents is often shared between high-level and low-level agents.

II. THE MORAL PROXY PROBLEM

Artificial agents are designed by humans to serve human ends and/or make decisions on their behalf, in areas where previously human agents would make decisions. They are, in the words of *Deborah Johnson* and *Keith Miller* 'tethered to humans'.[5] At least as long as artificial agents are not advanced enough to merit the ascription of moral responsibility in their own right, we can think of them as 'moral proxies' for human agents,[6] that is, as an extension of the agency of the humans on whose behalf they are acting. In any given context, the question then arises who they should be moral proxies for. I will refer to the problem of determining who, in any particular context, artificial agents ought to be moral proxies for as the 'Moral Proxy Problem'. This problem has been raised in different forms in a number of debates surrounding the design, ethics, politics, and legal treatment of artificial agents.

Take, for instance, the debate on the ethics of self-driving cars, where *Sven Nyholm* points out that before we apply various moral theories to questions of, for example, crash optimisation, we must settle on who the relevant moral agent is.[7] In the debate on value alignment – how to make sure the values advanced AI is pursuing are aligned with those of humans[8] – the Moral Proxy Problem arises as the question of whose values AI ought to be aligned with, especially in the context of reasonable disagreement between various stakeholders.[9] In computer science, *Vincent Conitzer* has recently raised the question of 'identity design', that is, the question of where one artificial agent ends and another begins.[10] He claims that how we should approach identity design depends at least in part on whether we want to be able to assign separate artificial agents to each user, so that they can represent their users separately, or are content with larger agents that can presumably only be understood as moral proxies for larger collectives of human agents. Finally, in debates around moral responsibility and legal liability for potential harms caused by artificial agents, the Moral Proxy Problem arises in the context of the question of which human agent(s) can be held responsible and accountable when artificial agents are not proper bearers of responsibility themselves.

For the purposes of my argument, I would like to distinguish between two types of answers to the Moral Proxy Problem: on the one hand, we could think of artificial agents as moral proxies for what I will call 'low-level agents', by which I mean the types of agents who would have faced the individual choice scenarios now faced by artificial agents in their absence, for example, the individual users of artificial agents such as owners of self-driving cars, or local authorities using artificial health decision systems. On the other hand, we could think of them as moral proxies for

[5] DG Johnson and KW Miller, 'Un-Making Artificial Moral Agents' (2008) 10 *Ethics and Information Technology* 123.
[6] See J Millar, 'Technology as Moral Proxy Autonomy and Paternalism by Design' (2015) 34 *IEEE Technology and Society Magazine* 47. Also see K Ludwig, 'Proxy Agency in Collective Action' (2014) 48 *Noûs* 75 for a recent analysis of proxy agency, J Himmelreich, 'Agency and Embodiment: Groups, Human–Machine Interactions, and Virtual Realities' (2018) 31 *Ratio* 197 on proxy agency as disembodied agency and S Köhler, 'Instrumental Robots' (2020) 26 *Science and Engineering Ethics* 3121 on artificial agents as 'instruments' for human agents.
[7] S Nyholm, 'The Ethics of Crashes with Self-Driving Cars: A Roadmap, I' (2018) 13(7) *Philosophy Compass* 6.
[8] See, e.g. S Russell, *Human Compatible: AI and the Problem of Control* (2019) for a prominent book-length treatment.
[9] See, e.g. I Gabriel, 'Artificial Intelligence, Values and Alignment' (2020) 30 *Minds and Machines* 411 for discussion.
[10] V Conitzer, 'Designing Preferences, Beliefs, and Identities for Artificial Intelligence' (2020) 33(1) *Proceedings of the AAI Conference on Artificial Intelligence (hereafter Conitzer, 'Designing Preferences')*.

what I will call 'high-level agents', by which I mean those who are in a position to potentially control the choice behaviour of many artificial agents,[11] such as designers of artificial agents, or regulators representing society at large.

I would also like to distinguish between two broad and connected purposes for which an answer to the Moral Proxy Problem is important, namely, ascription of responsibility and accountability on the one hand, and decision structuring for the purposes of design choices on the other. To start with the first purpose, here we are interested in who can be held responsible, in a backward-looking sense, for harms caused by artificial agents, which might lead to residual obligations, for example, to compensate for losses, but also who, in a forward-looking sense, is responsible for oversight and control of artificial agents. It seems natural that in many contexts, at least a large part of both the backward-looking and forward-looking responsibility for the choices made by artificial agents falls on those human agents whose moral proxies they are.

My primary interest in this paper is not the question of responsibility ascription, however, but rather the question of decision structuring, that is, the question of how the decision problems faced by artificial agents should be framed for the purposes of making design choices. The question of who is the relevant agent is in a particular context is often neglected in decision theory and moral philosophy but is crucial in particular for determining the scope of the decision problem to be analysed.[12] When we take artificial agents to be moral proxies for low-level human agents, it is natural to frame the relevant decisions to be made by artificial agents from the perspective of the low-level human agent. For instance, we could consider various problematic driving scenarios a self-driving car might find itself in, and then discuss how the car should confront these problems on behalf of the driver. Call this 'low-level agential framing'. When we take artificial agents to be moral proxies for high-level agents, on the other hand, we should frame the relevant decisions to be made by artificial agents from the perspective of those high-level agents. To use the example of self-driving cars again, from the perspective of designers or regulators, we should consider the aggregate consequences of many self-driving cars repeatedly confronting various problematic driving scenarios in accordance with their programming. Call this 'high-level agential framing'.

The issues of responsibility ascription and decision structuring are of course connected: when it is appropriate to frame a decision problem from the perspective of a particular agent, this is usually because the choice to be made falls under that agent's responsibility. Those who think of artificial agents as moral proxies for low-level agents often argue in favour of a greater degree of control on the part of individual users, for instance by having personalisable ethics settings, whereby the users can alter their artificial agent's programming to more closely match their own moral views.[13] Given such control, both decision structuring as well as most of the responsibility for the resulting choices should be low-level. But it is important to note here that the appropriate level of agential framing of the relevant decision problems and the level of agency at which we ascribe responsibility may in principle be different. We could, for instance, think of designers of

[11] Or, depending on your views on proper 'identity design' (see Conitzer, 'Designing Preferences' (n 10)) one single artificial agent making decisions in many decision contexts previously faced by many humans (e.g. a network of artificial trading agents acting in coordinated ways).

[12] See SO Hansson, 'Scopes, Options, and Horizons: Key Issues in Decision Structuring' (2018) 21 *Ethical Theory and Moral Practice* 259 for a very instructive discussion of this and other issues in decision structuring.

[13] See, e.g. A Sandberg and H Bradshaw-Martin, 'Autonomous Cars and their Moral Implications' (2015) 58(1) *Multitudes* 62; and G Contissa, F Lagioia, and G Sartor, 'The Ethical Knob: Ethically-Customisable Automated Vehicles and the Law' (2017) 25 *Artificial Intelligence and Law* 365.

artificial agents doing their best to design artificial agents to act on behalf of their users, but without giving the users any actual control over the design. As such, the designers could try to align the artificial agents with their best estimate of the users' considered and informed values. In that case, decision framing should be low-level. But insofar as low-level agents aren't actually in control of the programming of the artificial agents, we might think their responsibility for the resulting choices is diminished and should still lie mostly with the designers.

How should we respond to the Moral Proxy Problem for the purposes of decision structuring, then? In the literature on ethical dilemmas faced by artificial agents, a low-level response is often presupposed. The presumption of many authors there is that we can conclude fairly directly from moral judgements about individual dilemma situations (e.g., the much discussed trolley problem analogues) to how the artificial agents employed in the relevant context should handle them.[14] There is even an empirical ethics approach to making design decisions, whereby typical responses to ethical dilemmas that artificial agents might face are crowd-sourced, and then used to inform design choices.[15] This reflects an implied acceptance of artificial agents as low-level moral proxies. The authors mentioned who are arguing in favour of personalisable ethics settings for artificial agents also appear to be presupposing that the artificial agents they have in mind are moral proxies for low-level agents. The standard case for personalisable ethics settings is based on the idea that mandatory ethics settings would be unacceptably paternalistic. But imposing a certain choice on a person is only paternalistic if that choice was in the legitimate sphere of agency of that person in the first place. Saying that mandatory ethics settings are paternalistic thus presupposes that the artificial agents under discussion are moral proxies for low-level agents.

What could be a positive argument in favour of low-level agential framing? I can think of two main ones. The first draws on the debate over responsibility ascription. Suppose we thought that, in some specific context, the only plausible way of avoiding what are sometimes called 'responsibility gaps', that is, of avoiding cases where nobody can be held responsible for harms caused by artificial agents, was to hold low-level agents, and in particular users, responsible.[16] Now there seems to be something unfair about holding users responsible for choices by an artificial agent that (a) they had no design control over, and that (b) are only justifiable when framing the choices from a high-level agential perspective. Provided that, if we were to frame choices from a high-level agential perspective, we may sometimes end up with choices that are not justifiable from a low-level perspective, this provides us with an argument in favour of low-level agential framing. Crucially, however, this argument relies on the assumption that only low-level agents can plausibly be held responsible for the actions of artificial agents, which is of course contested, as well as on the assumption that there is sometimes a difference between what is morally justifiable when adopting a high-level and a low-level agential framing respectively, which I will return to.

[14] See, e.g. P Lin, 'Why Ethics Matter for Autonomous Cars' in M Maurer and others (eds), *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte* (2015); G Keeling, 'Why Trolley Problems Matter for the Ethics of Automated Vehicles' (2020) 26 *Science and Engineering Ethics* 293 (hereafter Keeling, 'Trolley Problems'). See also J Himmelreich, 'Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations' (2018) 21 *Ethical Theory and Moral Practice* 669 (hereafter Himmelreich, 'Ethics of Autonomous Vehicles'); Nyholm and Smids, 'Ethics of Accident-Algorithms' (n 4); and A Jaques, 'Why the Moral Machine Is a Monster' (2019) University of Miami Law School: We Robot Conference (hereafter Jaques, 'Why the Moral Machine Is a Monster') who find issue with this.

[15] See E Awad and others, 'Crowdsourcing Moral Machines' (2020) 63(3) *Communications of the ACM* 48.

[16] On this and other solutions to the threat of responsibility gaps in the legal context see S Beck, 'The Problem of Ascribing Legal Responsibility in the Case of Robotics' (2016) 31 *AI Society* 473 (hereafter Beck, 'Ascribing Legal Personality'). For instance, German law assigns liability for damage caused by parking distance control systems to individual users.

A second potential argument in favour of a low-level response to the Moral Proxy Problem is based on the ideal of liberal neutrality, which is indeed sometimes invoked to justify anti-paternalism of the form proponents of personalisable ethics settings are committed to. The moral trade-offs we can expect many artificial agents to face are often ones there is reasonable disagreement about. We are then, in Rawlsian terms, faced with a political, not a moral problem:[17] how do we ensure fair treatment of all given reasonable pluralism? In such contexts, one might think higher-level agents, such as policy-makers or tech companies should maintain liberal neutrality; they should not impose one particular view on an issue that reasonable people disagree on. One way of maintaining such neutrality in the face of a plurality of opinion is to partition the moral space so that individuals get to make certain decisions themselves.[18] In the case of artificial agents, such a partition of the moral space can be implemented, it seems, by use of personalisable ethics settings, which implies viewing artificial agents as moral proxies for low-level agents.

At the same time, we also find in the responses to arguments in favour of personalisable ethics settings some reasons to think that perhaps there is not really much of a conflict, in practice, between taking a high-level and a low-level agential perspective. For one, in many potential contexts of application of artificial agents, there are likely to be benefits from coordination between artificial agents that each individual user can in fact appreciate. For instance, *Jan Gogoll* and *Julian Müller* point out the potential for collective action problems when ethics settings in self-driving cars are personalisable: each may end up choosing a 'selfish' setting, even though everybody would prefer a situation where everybody chose a more 'altruistic' setting.[19] If that is so, it is in fact in the interest of everybody to agree to a mandatory 'altruistic' ethics setting. Another potentially more consequential collective action problem in the case of self-driving cars is the tragedy of the commons when it comes to limiting emissions, which could be resolved by mandatory programming for fuel-efficient driving. And *Jason Borenstein*, *Joseph Herkert*, and *Keith Miller* point out the advantages, in general, of a 'systems-level analysis', taking into account how different artificial agents interact with each other, as their interactions may make an important difference to outcomes.[20] For instance, a coordinated driving style between self-driving cars may help prevent traffic jams and thus benefit everybody.

What this points to is that in cases where the outcomes of the choices of one artificial agent depend on what everybody else does and vice versa, and there are potential benefits for each from coordination and cooperation, it may seem like there will not be much difference between taking a low-level and a high-level agential perspective. From a low-level perspective, it makes sense to agree to not simply decide oneself how one would like one's artificial agent to choose. Rather, it is reasonable from a low-level perspective to endorse a coordinated choice where designers or regulators select a standardised programming that is preferable for each individual compared to the outcome of uncoordinated choice. And notably, this move does not need to be in tension with the ideal of liberal neutrality either: in fact, finding common principles that can be endorsed from each reasonable perspective is another classic way to ensure liberal neutrality

[17] As also pointed out by Himmelreich 'Ethics of Autonomous Vehicles' (n 14) and I Gabriel, 'Artificial Intelligence, Values and Alignment' (2020) 30 *Minds and Machines* 411.

[18] See J Gogoll and J Müller, 'Autonomous Cars: In Favor of a Mandatory Ethics Setting' (2017) 23 *Science and Engineering Ethics* 681 (hereafter Gogoll and Müller, 'Autonomous Cars') for this proposal, though they ultimately reject it. The phrase 'partition of the moral space' is due to G Gaus, 'Recognized Rights as Devices of Public Reason', in J Hawthorne (ed), *Ethics, Philosophical Perspectives* (2009), 119.

[19] Ibid.

[20] J Borenstein, J Herkert and K Miller, 'Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis' (2019) 25 *Science and Engineering Ethics* 383. See also Jaques, 'Why the Moral Machine Is a Monster' (n 14).

in the face of reasonable pluralism, in cases where partitioning the moral space in the way previously suggested can be expected to be worse for all. In the end, the outcome may not be so different from what a benevolent or democratically constrained high-level agent would have chosen if we thought of the artificial agents in question as high-level proxies in the first place.

Another potential reason for thinking that there may not really be much of a conflict between taking a high-level and a low-level agential perspective appears plausible in the remaining class of cases where we don't expect there to be much of an interaction between the choices of one artificial agent and any others. And that is simply the thought that in such cases, what a morally reasonable response to some choice scenario is should not depend on agential perspective. For instance, one might think that what a morally reasonable response to some trolley-like choice scenario is should not depend on whether we think of it from the perspective of a single low-level agent, or as part of a large number of similar cases a high-level agent is deciding on.[21] And if that is so, at least for the purposes of decision structuring, it would not make a difference whether we adopt a high-level or a low-level agential perspective. Moreover, the first argument we just gave in favour of low-level agential framing would be undercut.

Of course, while this may result in the Moral Proxy Problem being unimportant for the purposes of decision structuring, this does not solve the question of responsibility ascription. Resolving that question is not my primary focus here. What I would like to point out, however, is that the idea that agential framing is irrelevant for practical purposes sits nicely with a popular view on the question of responsibility ascription, namely the view that responsibility is often distributed among a variety of agents, including both high-level and low-level agents. Take, for instance, *Mariarosaria Taddeo* and *Luciano Floridi*:

> The effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware [...] With distributed agency comes distributed responsibility.[22]

Shared responsibility between, amongst others, designers and users is also part of Rule 1 of 'the Rules' for moral responsibility of computing artefacts championed by *Miller*.[23] The reason why the idea of shared responsibility sits nicely with the claim that agential framing is ultimately practically irrelevant is that in that case, no agent can be absolved from responsibility on the grounds that whatever design choice was made was not justifiable from their agential perspective.The following discussion will put pressure on this position. It will show that in the context of risk, quite generally, agential perspective in decision structuring is practically relevant. This is problematic for the view that responsibility for the choices of artificial agents is often shared between high-level and low-level agents and puts renewed pressure on us to address the Moral Proxy Problem in a principled way. I will return to the Moral Proxy Problem in Section V to discuss why this is, in fact, a hard problem to address. In particular, it will become apparent that

[21] A Wolkenstein, 'What has the Trolley Dilemma Ever Done for Us (and What Will it Do in the Future)? On some Recent Debates about the Ethics of Self-Driving Cars' (2018) 20 *Ethics and Information Technology* 163 seems to make essentially this claim in response to criticism of the importance of trolley cases when thinking of the ethics of self-driving cars.

[22] M Taddeo and L Floridi, 'How AI Can Be a Force for Good' (2018) 361 *Science* 751, 751. See also M Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 *Science and Engineering* 2051. In the legal literature, there has been appeal to the idea of a legal 'electronic person' composed of designers, producers, users, etc. as a potential responsibility-bearer, see for example Beck, 'Ascribing Legal Personality' (n 16).

[23] K Miller, 'Moral Responsibility for Computing Artifacts: "The Rules"' (2011) 13(3) *IT Professional* 57.

the low-level response to the problem that is so commonly assumed comes with significant costs in many applications. The high-level alternative, however, is not unproblematic either.

III. THE LOW-LEVEL CHALLENGE TO RISK NEUTRALITY IN ARTIFICIAL AGENT DESIGN

I now turn to the question of how to design artificial agents to deal with risk, which I will go on to argue is a practical design issue which crucially depends on our response to the Moral Proxy Problem. Expected utility theory is the orthodox theory of rational choice under conditions of risk that, on the standard approach, designers of artificial agents eventually aim to implement. The theory is indeed also accepted by many social scientists and philosophers as a theory of instrumentally rational choice, and moreover incorporated by many moral philosophers when theorising about our moral obligations in the context of risk.[24] Informally, according to this theory, for any agent, we can assign both a probability and a utility value to each potential outcome of the choices open to them. We then calculate, for each potential choice, the probability-weighted sum of the utilities of the different potential outcomes of that choice. Agents should make a choice that maximises this probability-weighted sum.

One widely discussed worry about applying expected utility theory in the context of artificial agent design is that when risks of harm are imposed on others, the application of expected utility theory implies insensitivity to how ex ante risks are distributed among the affected individuals.[25] For instance, suppose that harm to one of two individuals is unavoidable, and we judge the outcomes where one or the other is harmed to be equally bad. Expected utility theory then appears unable to distinguish between letting the harm occur for certain for one of the individuals, and throwing a fair coin, which would give each an equal chance of being harmed. Yet the latter seems like an intuitively fairer course of action.

In the following, I would like to abstract away as much as possible from this problem, but rather engage with an independent concern regarding the use of expected utility theory when designing artificial agents that impose risks on others. And that is that, at least under the interpretation generally adopted for artificial agent design, the theory implies risk neutrality in the pursuit of goals and values, and rules out what we will call 'pure' risk aversion (or pure risk seeking), as I will explain in what follows. Roughly, risk aversion in the attainment of some good manifests in settling for an option with a lower expectation of that good because the range of potential outcomes is less spread out, and there is thus a lesser risk of ending up with bad outcomes. For instance, choosing a certain win of £100 over a 50% chance of £300 would be a paradigmatic example of risk aversion with regard to money. The expected monetary value of the 50% gamble is £150. Yet, to the risk averse agent, the certain win of £100 may be preferable because the option does not run the risk of ending up with nothing.

Expected utility theory can capture risk aversion through decreasing marginal utility in the good. When marginal utility is decreasing for a good, that means that, the more an agent already has of a good, the less additional utility is assigned to the next unit of the good. In our example, decreasing marginal utility may make it the case that the additional utility gained from receiving

[24] Indeed, as remarks by Keeling exemplify in the case of this debate, moral philosophers often assume that there can be a division of labour between them and decision theorists, whereby issues to do with risk and uncertainty are settled by decision theorists alone. For more see Keeling, 'Trolley Problems' (n 14). The issues discussed in the following illustrate just one way in which this assumption is mistaken.

[25] On the general issue of fair risk imposition, see the useful overview by M Hayenhjelm and J Wolff, 'The Moral Problem of Risk Imposition: A Survey of the Literature' (2012) 20 *European Journal of Philosophy* 26.

£100 is larger than the additional utility gained from moving from £100 to £300. If that is so, then the risk averse preferences we just described can be accommodated within expected utility theory: the expected utility of a certain £100 will be higher than the expected utility of a 50% chance of £300 – even though the latter has higher expected monetary value.

Whether this allows us to capture all ordinary types of risk aversion depends in part on what we think utility is. According to what we might call a 'substantive' or 'realist' understanding of utility, utility is a cardinal measure of degrees of goal satisfaction or value. On that view, expected utility theory requires agents to maximise their expected degree of goal satisfaction, or expected value. And having decreasing marginal utility, on this view, means that the more one already has of a good, the less one values the next unit, or the less the next unit advances one's goals. On this interpretation, only agents who have decreasing marginal utility in that sense are permitted to be risk averse within expected utility theory. What is ruled out is being risk averse beyond what is explainable by the decreasing marginal value of a good. Formally, expected utility theory does not allow agents to be risk averse with regard to utility itself. On this interpretation, that means agents cannot be risk averse with regard to degrees of goal satisfaction, or value itself, which is what the above reference to 'pure' risk aversion is meant to capture. For instance, on this interpretation of utility, expected utility theory rules out that an agent is risk averse despite valuing each unit of a good equally.[26]

Importantly for us, such a substantive conception of utility seems to be widely presupposed both in the literature on the design of artificial agents, as well as by those moral philosophers who incorporate expected utility theory when thinking about moral choice under risk. In moral philosophy, expected utility maximisation is often equated with expected value maximisation, which, as we just noted, implies risk neutrality with regard to value itself.[27] When it comes to artificial agent design, speaking in very broad strokes, on the standard approach we start by specifying the goals the system should be designed to pursue in what is called the 'objective function' (or alternatively, the 'evaluation function', 'performance measure', or 'merit function'). For very simple systems, the objective function may simply specify one goal. For instance, we can imagine an artificial nutritional assistant whose purpose it is simply to maximise caloric intake. But in most applications, the objective function will specify several goals, as well how they are to be traded off. For instance, the objective function for a self-driving car will specify that it should reach its destination fast; use little fuel; avoid accidents and minimise harm in cases of unavoidable accident; and make any unavoidable trade-offs between these goals in a way that reflects their relative importance.

[26] On this and other interpretations of utility, see J Thoma, 'Decision Theory' in R Pettigrew and J Weisberg (eds), *The Open Handbook of Formal Epistemology* (2019). Note that there is a way of understanding utility that is popular amongst economists which does not have that implication. On what we may call the 'formal' or 'constructivist' interpretation, utility is merely whatever measure represents an agent's preferences, provided these preferences are consistent with the axioms of a representation theorem for the version of expected utility theory one is advocating. According to that understanding, what expected utility theory requires of agents is having consistent preferences, so that they are representable as expected utility maximising. And in that case, having decreasing marginal utility just expresses the fact that one is risk averse, because that must be a feature of the agent's utility function if we are to capture her as risk averse and expected utility maximising. Importantly, on this view, because utility is not assumed to be a cardinal measure of value itself, we can allow for the utility function to exhibit decreasing marginal utility in value or degrees of goal satisfaction, thus allowing for pure risk aversion.

[27] Specifically in the debate about the ethics of artificial agents, this assumption is made, for example by A Hevelke and J Nida-Rümelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21 *Science and Engineering Ethics* 619; N Goodall, 'Away from Trolley Problems and toward Risk Management' (2016) 30 *Applied Artificial Intelligence* 820; Gogoll and Müller, 'Autonomous Cars' (n 17); and Keeling, 'Trolley Problems' (n 14) among many others.

After we have specified the objective function, the artificial agent should be either explicitly programmed or trained to maximise the expectation of that objective function.[28] Take, for instance, this definition of rationality from *Stuart Russell* and *Peter Norvig's* textbook on AI:

> For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.[29]

According to the authors, the goal of artificial agent design is to implement this notion of rationality as well as possible. But this just means implementing expected utility theory under a substantive understanding of the utility function as a performance measure capturing degrees of goal satisfaction.[30]

So, we have seen that, by assuming a substantive interpretation of utility as a cardinal measure of value or degrees of goal satisfaction, many moral philosophers and designers of artificial agents are committed to risk neutrality with regard to value or goal satisfaction itself. However, such risk neutrality is not a self-evident requirement of rationality and/or morality. Indeed, some moral philosophers have defended a requirement to be risk averse, for instance when defending precautionary principles of various forms, or famously *John Rawls* in his treatment of choice behind the veil of ignorance.[31] And the risk neutrality of expected utility theory under the substantive interpretation of utility has been under attack recently in decision theory as well, for example by *Lara Buchak*.[32]

To illustrate, let me introduce two scenarios that an artificial agent might find itself in, where the risk neutral choice appears intuitively neither morally nor rationally required, and where indeed many human agents can be expected to choose in a risk averse manner.

Case 1: Artificial Rescue Coordination Centre. An artificial rescue coordination centre has to decide between sending a rescue team to one of two fatal accidents involving several victims. If it chooses Accident 1, one person will be saved for certain. If it chooses Accident 2, on the other hand, there is a 50% chance of saving three and a 50% chance of saving nobody. It seems plausible in this case that the objective function should be linear in lives saved, all other things being equal – capturing the idea that all lives are equally valuable. And let us suppose that all other morally

[28] On the difference between top-down and bottom-up approaches to implementing ethical design, see C Allen, I Smit, and W Wallach, 'Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches' (2005) 7 *Ethics and Information Technology* 149. What is important for us is the ideal of maximising the expectation of the objective function that is shared by both.

[29] S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (4th ed., 2020) 37.

[30] It should be noted here that these authors also speak explicitly of 'utility' as distinct from the performance measure, because they think of the utility function more narrowly as something that is used internally by an agent to compute an optimal choice. For those agents that are programmed to be such explicit expected utility maximisers, the authors do remark that 'an agent's utility function is essentially an internalization of the performance measure. If the internal utility function and the external performance measure are in agreement, then an agent that chooses actions to maximize its utility will be rational according to the external performance measure.' (Ibid, 53) But note that expected utility theory is not normally understood to require explicit deliberation involving a utility function, but to also accommodate agents whose choices de facto maximise expected utility, no matter what deliberative procedure they use. Russell and Norvig's definition of rationality captures this wider conception of expected utility theory if we think of utility and performance measure as equivalent.

[31] J Rawls, *A Theory of Justice* (1971).

[32] L Buchak, *Risk and Rationality* (2013) (hereafter Buchak, *Risk and Rationality*).

relevant factors are indeed equal between the two options open to the rescue coordination centre.[33] In this scenario, a risk neutral rescue coordination centre would always choose Accident 2, because the expected number of lives saved (1.5) is higher. However, I submit that many human agents, if they were placed in this situation with time to deliberate, would choose Accident 1 and thus exhibit risk aversion. Moreover, doing so is neither intuitively irrational nor immoral. If this is not compelling, consider the case where attending to Accident 2 comes with only a 34% chance of saving three. Risk neutrality still requires choosing Accident 2. But it is very hard to see what would be morally or rationally wrong with attending to Accident 1 and saving one for certain instead.

Case 2: Changing Lanes. A self-driving car is driving in the left lane of a dual carriageway in the UK and is approaching a stumbling person on the side of the road. At the same time, a car with two passengers is approaching from behind on the right lane. The self-driving car estimates there is a small chance the other car is approaching fast enough to fatally crash into it should it change lanes (Changing Lanes), and a small albeit three times higher chance that the person on the side of the road could trip at the wrong time and consequently be fatally hit by the self-driving car should it not change lanes (Not Changing Lanes). Specifically, suppose that Not Changing Lanes comes with a 0.3% chance of killing one, meaning the expected number of fatalities is 0.003. Changing Lanes, on the other hand, comes with a 0.1% chance of killing two, meaning the expected number of fatalities is 0.002. Suppose that the passenger of the self-driving car will be safe either way. It seems plausible that the objective function should be linear in accidental killings, all other things being equal – again capturing the idea that all lives are equally valuable. And let us again suppose that all other morally relevant factors are indeed equal between the two options open to the self-driving car. In this scenario, a risk neutral car would always choose Changing Lanes, because the expected number of fatalities is lower. However, I submit that many human agents would, even with time to reflect, choose Not Changing Lanes to rule out the possibility of killing 2, and thus exhibit risk aversion. Moreover, doing so is neither intuitively irrational nor immoral.

These admittedly very stylised cases were chosen because they feature an objective function uncontroversially linear in the one value at stake, in order to illustrate the intuitive permissibility of pure risk aversion. Most applications will, of course, feature more complex objective functions trading off various concerns. In such cases, too, the standard approach to artificial agent design requires risk neutrality with regard to the objective function itself. But again, it is not clear why risk aversion should be ruled out, for example when a nursebot that takes into account both the potential value of saving a life and the cost of calling a human nurse faces a risky choice about whether to raise an alarm.

In the case of human agents, we tend to be permissive of a range of pure risk attitudes, including different levels of pure risk aversion. There appears to be rational and moral leeway on degrees of risk aversion, and thus room for reasonable disagreement. Alternatives to expected utility theory, such as Buchak's risk-weighted expected utility theory, as well as expected utility theory under some interpretations other than the substantive one, can accommodate such

[33] Note in particular that, if we don't allow randomisation between the two options, ex ante equality is impossible to achieve, and thus not a relevant factor.

rational leeway on attitudes towards risk.[34] But the commitment to expected utility theory under a substantive interpretation of utility, as we find it in the literature on the design of artificial agents, rules this out and imposes risk neutrality instead – which is a point not often acknowledged, and worth emphasising.

To return to the Moral Proxy Problem, suppose that we want artificial agents to be low-level moral proxies. In the preceding examples, we have already picked the right agential framing then: we have structured the relevant decision problem as an individual choice situation as it might previously have been faced by a low-level human agent. A low-level moral proxy should, in some relevant sense, choose in such a way as to implement considered human judgement from the low-level perspective. Under risk, this plausibly implies that we should attempt to align not only the artificial agent's evaluations of outcomes, but also its treatment of risk to the values and attitudes of the low-level agents it is a moral proxy for. There are different ways of making sense of this idea, but on any such way, it seems like we need to allow for artificial agents to sometimes exhibit risk aversion in low-level choices like the ones just discussed.

As we have seen before, some authors who view artificial agents as low-level moral proxies have argued in favour of personalisable ethics settings. If there is, as we argued, reasonable disagreement about risk attitudes, artificial agents should then also come with personalisable risk settings. If we take an empirical approach and crowd-source and then implement typical judgements on ethical dilemma situations like the ones just discussed, we will likely sometimes need to implement risk averse judgements as well. Lastly, in the absence of personalisable ethics and risk settings but while maintaining the view of artificial agents as low-level moral proxies, we can also view the design decision as the problem of how to make risky choices on behalf of another agent while ignorant of their risk attitudes. One attractive principle for how to do so is to implement the most risk averse of the reasonable attitudes towards risk, thereby erring on the side of being safe rather than sorry when choosing for another person.[35] Again, the consequence would be designing artificial agents that are risk averse in low-level decisions like the ones we just considered.

We have seen, then, that in conflict with the standard approach to risk in artificial agent design, if we take artificial agents to be low-level moral proxies, we need to allow for them to display pure risk aversion in some low-level choice contexts like the ones just considered. The next section will argue that things look quite different, however, if we take artificial agents to be high-level moral proxies.

IV. RISK AVERSION AND THE HIGH-LEVEL AGENTIAL PERSPECTIVE

Less stylised versions of the scenarios we just looked at are currently faced repeatedly by different human agents and will in the future be faced repeatedly by artificial agents. While such decisions are still made by human agents, there is usually nobody who is in control of a large number such choice problems: human rescue coordinators will usually not face such a dramatic decision multiple times in their lives. And most drivers will not find themselves in such dangerous driving situations often. The regulatory reach of higher-order agents such as policy-makers over human agents is also likely to be limited in these scenarios and many other areas in which artificial agents might be introduced to make decisions in place of humans – both because human agents in such choice situations have little time to reflect and will thus often

[34] Buchak, *Risk and Rationality* (n 32).

[35] For a defence of such a principle see L Buchak, 'Taking Risks Behind the Veil of Ignorance' (2017) 127(3) *Ethics* 610.

be excused for not following guidelines, and because, in the case of driving decisions in particular, there are limits to the extent to which drivers would accept being micromanaged by the state.

Things are different, however, once artificial agents are introduced. Now there are higher-level agents, in particular designers, who can directly control the choice behaviour of many artificial agents in many instances of the decision problems we looked at in the last section. Moreover, these designers have time to reflect on how decisions are to be made in these choice scenarios and have to be explicit about their design choice. This also gives greater room for other higher-level agents, such as policy-makers, to exert indirect control over the choices of artificial agents, by regulating the design of artificial agents. Suppose we think that artificial agents in some specific context should in fact be thought of as moral proxies not for low-level agents such as individual users of self-driving cars, but rather as moral proxies for such high-level agents. From the perspective of these higher-level agents, what seems most relevant for the design choice are the expected aggregate consequences of designing a whole range of artificial agents to choose in the specified ways on many different occasions. I want to show here that this makes an important difference in the context of risk.

To illustrate, let us return to our stylised examples, starting with a modified version of Case 1: Artificial Rescue Coordination Centre:

> Suppose some high-level agent has to settle at once on one hundred instances of the choice between Accident 1 and Accident 2. Further, suppose these instances are probabilistically independent, and that the same choice needs to be implemented in each case. The two options are thus always going for Accident 1, saving one person for certain each time, or always going for Accident 2, with a 50% chance of saving three each time. The expected aggregate outcome of going for Accident 1 one hundred times is, of course, saving one hundred people for certain. The expected aggregate result of going for Accident 2 one hundred times, on the other hand, is a probability distribution with an expected number of one hundred and fifty lives saved, and, importantly, a <0.5% chance of saving fewer lives than if one always went for Accident 1. In this compound case, it now seems unreasonably risk averse to choose the 'safe option'.

Similarly, if we look at a compound version of Case 2: Changing Lanes:

> Suppose a higher-level agent has to settle at once how 100,000 instances of that choice should be made, where these are again assumed to be probabilistically independent, and the same choice has to be made on each instance. One could either always go for the 'safe' option of Not Changing Lanes. In that case, the expected number of fatalities is 300, with a <0.1% chance of less than 250 fatalities. Or one could always go for the 'risky' option of Changing Lanes. In that case, the expected number of fatalities is only 200, with only a ~0.7% chance of more than 250 fatalities. As before, the 'risky' option is thus virtually certain to bring about a better outcome in the aggregate, and it would appear unreasonably risk averse to stick with the 'safe' option.

In both cases, as the number of repetitions increases, the appeal of the 'risky' option only increases, because the probability of doing worse than on the 'safe' option becomes ever smaller. We can also construct analogous examples featuring more complex objective functions appropriate for more realistic cases. It remains true that as independent instances of the risky choice problem are repeated, at some point the likelihood of doing better by each time choosing a safer option with lower expected value becomes very small. From a sufficiently large compound perspective, the virtual certainty of doing better by picking a riskier option with higher expected value is decisive. And thus, when we think of artificial agents as moral proxies for high-level agents that are in a position to control sufficiently many low-level decisions, designing the

artificial agents to be substantially risk averse in low-level choices seems impermissible. From the high-level agential perspective, the risk neutrality implied by the current standard approach in artificial agent design seems to, in fact, be called for.[36]

The choice scenarios we looked at are similar to a case introduced by *Paul Samuelson*,[37] which I discuss in more detail in another paper.[38] *Samuelson's* main concern there is that being moderately risk averse in some individual choice contexts by, for example, choosing the safer Accident 1 or Not Changing Lanes, while at the same time choosing the 'risky' option in compound cases is not easily reconcilable with expected utility theory (under any interpretation).[39] It is undeniable, though, that such preference patterns are very common. And importantly, in the cases we are interested in here, no type of agent can actually be accused of inconsistency, because we are dealing with two types of agents with two types of associated choice problems. One type of agent, the low-level agent who is never faced with the compound choice, exhibits the reasonable seeming risk averse preferences regarding 'small-scale' choices to be made on her behalf. And another type of agent, the high-level agent, exhibits again reasonable-seeming preferences in compound choices that translate to effective risk neutrality in each individual 'small-scale' choice scenario.

The take-away is thus that how we respond to the Moral Proxy Problem is of practical relevance here: If we take artificial agents to be moral proxies for low-level agents, they will sometimes need to be programmed to exhibit risk aversion in the kinds of individual choice contexts where they are replacing human agents. If we take them to be moral proxies for high-level agents, they should be programmed to be risk neutral in such choice contexts, as the current approach to risk in artificial agent design in fact implies, because this has almost certainly better consequences in the aggregate.

V. BACK TO THE MORAL PROXY PROBLEM

We saw in Section II that the Moral Proxy Problem matters for decision structuring: whether we take artificial agents to be moral proxies for low-level or high-level agents determines from which agential perspective we are framing the relevant decision problems. I raised the possibility, alluded to by some authors, that resolving the Moral Proxy Problem one way or the other is of little practical relevance, because agential framing does not make a practical difference for design choices. The issue of whether artificial agents should be designed to be risk neutral or allowed to be risk averse, discussed in the last two sections, is then an especially challenging one in the context of the Moral Proxy Problem, because it shows the hope for this irrelevance to be ungrounded: agential perspective turns out to be practically crucial.

Notably, the stylised examples we discussed do not describe collective action or coordination problems where each can recognise from her low-level perspective that a higher-level agent could implement a coordinated response that would be superior from her perspective and everybody else's. Crucially, both the outcomes and the probabilities in each of the lower-level

[36] To the extent that even low-level agents face some risky decisions very often, we may also take this to be an argument that in those cases, risk neutrality in the individual choice instances is called for even from the low-level perspective. However, in our examples, the individual choice scenarios are both rare and high-stakes from the low-level perspective, so that the compound perspective really only becomes relevant for high-level agents. It is in that kind of context that agential perspective makes a crucial practical difference.

[37] P Samuelson, 'Risk and Uncertainty: A Fallacy of Large Numbers' (1963) 98 *Scientia* 108 (hereafter Samuelson, 'Risk and Uncertainty').

[38] J Thoma, 'Risk Aversion and the Long Run' (2019) 129(2) *Ethics* 230.

[39] Samuelson, 'Risk and Uncertainty' (n 37).

choice contexts are independent in our examples. And having a particular design imposed by a higher-level agent does not change the potential outcomes and probabilities of the choice problem faced by any particular artificial agent. It only changes the actual choice in that lower-level choice problem from a potentially risk averse to a risk neutral one. This is not something that a risk averse lower-level agent would endorse.

It thus becomes practically important to resolve the Moral Proxy Problem. And for the purposes of decision structuring, at least, it is not an option to appeal to the notion of distributed agency to claim that artificial agents are moral proxies for both low-level and high-level agents. Adopting one or the other agential perspective will sometimes call for different ways of framing the relevant decision problem, and we need to settle on one specification of the decision problem before we can address it. Where we imagine there being a negotiation between different stakeholders in order to arrive at a mutually agreeable result, the framing of the decision problem to be negotiated on will also need to be settled first. For decision structuring, at least, we need to settle on one agential perspective.

For reasons already alluded to, the fact that substantially different designs may be morally justified when decision problems are framed from the high-level or the low-level agential perspective is also problematic for ascribing shared responsibility for the choices made by artificial agents. If different programmings are plausible from the high-level and low-level perspective, it may seem unfair to hold high-level agents (partially) responsible for choices justified from the low-level perspective and vice versa. If, based on a low-level framing, we end up with a range of risk averse self-driving cars that cause almost certainly more deaths in the aggregate, there is something unfair about holding designers responsible for that aggregate result. And if, based on a high-level framing, we in turn end up with a range of risk neutral self-driving cars, which, in crash scenarios frequently save nobody when they could have saved some for sure, there is something unfair about holding individual users responsible for that tough call they would not have endorsed from their perspective.[40] At least, it seems like any agent who will be held responsible for some (set of) choices has some rightful claim for the decision problem to be framed from their agential perspective. But where agential perspective makes a practical difference not all such claims can be fulfilled.

Let us return now to the problem of decision structuring, where, for the reasons just mentioned, we certainly need to resolve the Moral Proxy Problem one way or the other. However we resolve it, there are major trade-offs involved. I already mentioned some potential arguments in favour of low-level agential framing. There is, for one, the idea that low-level agential framing is natural if we want to hold low-level agents responsible. If we don't have an interest in holding low-level agents responsible, this is, of course, not a relevant consideration. But I would also like to add an observation about moral phenomenology that may have at least some political relevance. Note that users and owners of artificial agents are in various senses morally closer to the potentially harmful effects of the actions of their artificial agents than designers or policy-makers: they make the final decision of whether to deploy the agent; their lives may also be at stake; they often more closely observe the potentially harmful event and have to live with its memory; and users are often asked to generally maintain responsible oversight of the operations of the artificial agent. All this may, at least, result in them feeling more responsible for the actions of their artificial agent. Such a

[40] Granted, individual users usually do make the final call of whether to deploy an artificial agent and may do so knowing how they would act in certain morally difficult situations. Still, if certain aspects of the programming of the artificial agent one deploys only make sense from the perspective of general public safety, or general public health, and only in the context of many other artificial agents being programmed in the same way, it is natural to resist individual responsibility for the consequences of that aspect of the artificial agent's design.

feeling of responsibility and moral closeness without control, or without at least the sense that efforts were made for the choices of the artificial agent to capture one's considered judgements as well as possible is a considerable burden.

A second argument we made in favour of low-level agential framing appealed to the idea of liberal neutrality in the face of reasonable disagreement, which could be implemented effectively by partitioning the moral space so as to leave certain decisions up to individuals. Such partitioning seems like an effective way to implement liberal neutrality especially in the absence of collective action problems that may create general agreement on a coordinated response. Given the independence in outcomes and probabilities, the cases we have discussed indeed do not constitute such collective action problems, but they do feature reasonable disagreement in the face of rational and moral leeway about risk attitudes. I believe that the ideal of liberal neutrality is thus a promising consideration in favour of low-level agential framing.

What the preceding sections have also made clear, however, is that low-level agential framing in the context of risk may come at the cost of aggregate outcomes that are almost certainly worse than the expected consequences of the choices that seem reasonable from the high-level agential perspective. This consequence of low-level agential framing is, as far as I know, unacknowledged, and may be difficult for proponents of low-level agential framing to accept.

If we respond to the Moral Proxy Problem by adopting a high-level agential perspective in those contexts instead, this problem is avoided. And other considerations speak in favour of thinking of artificial agents as moral proxies for high-level agents. An intuitive thought is this: as a matter of fact, decisions that programmers and those regulating them make determine many lower-level choices. In that sense they are facing the compound choice, in which the almost certainly worse aggregate outcome of allowing lower-level risk aversion appears decisive. In order to design artificial agents to be (risk averse) moral proxies for individual users, designers would have to abstract away from these very real aggregate implications of their design decisions. This may put designers in a similarly difficult position to the owner of a self-driving car that she knows may make choices that seem reckless from her perspective.

Following on from this, arguments in favour of holding high-level agents responsible will also, at least to some extent, speak in favour of high-level agential framing, because again it seems high-level agential framing is natural when we want to hold high-level agents responsible. We find one potential argument in favour of ascribing responsibility to high-level agents in *Hevelke* and *Nida-Rümelin's* appeal to moral luck.[41] Their starting point is that whether individual artificial agents ever find themselves in situations where they have to cause harm is in part down to luck. For instance, it is in part a matter of luck whether, and if so how often, any artificial agent finds itself in a dangerous driving situation akin to the one described in Case 2 mentioned earlier. And, no matter how the agent chooses, it is a further matter of luck whether harm is actually caused. Where harm is caused, it may seem unfair to hold the unlucky users of those cars responsible, but not others who employed their artificial agents no differently. *Alexander Hevelke* and *Julian Nida-Rümelin* take this observation to speak in favour of ascribing responsibility collectively to the group of all users of a type of artificial agent. But finding responsibility with other high-level agents, such as the companies selling the artificial agents would also avoid the problem of moral luck. And then it also makes sense to adopt a high-level perspective for the purposes of decision structuring.

[41] A Hevelke and J Nida-Rümelin, 'Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis' (2015) 21 *Science and Engineering Ethics* 619.

Still, the practical relevance of agential framing also brings about and highlights costs of settling for a high-level solution to the Moral Proxy Problem that are worth stressing: this solution will mean that, where artificial agents operate in areas where previously human agents made choices, these artificial agents will make some choices that are at odds with even considered human judgement. And the high-level solution will introduce higher-level control, be it by governments or tech companies, in areas where previously decision-making by humans has been decentralised, and in ways that don't simply reproduce what individual human agents would have (ideally) chosen for themselves. In this sense, the high-level solution involves a significant restructuring of our moral landscape.

VI. CONCLUSION

I have argued that the Moral Proxy Problem, the problem of determining what level of human (group) agent artificial agents ought to be moral proxies for, has special practical relevance in the context of risk. Moral proxies for low-level agents may need to be risk averse in the individual choices they face. Moral proxies for high-level agents, on the other hand, should be risk neutral in individual choices, because this has almost certainly better outcomes in the aggregate. This has a number of important implications. For one, it means we actually need to settle, in any given context, on one response to the Moral Proxy Problem for purposes of decision structuring at least, as we don't get the same recommendations under different agential frames. This, in turn, puts pressure on the position that responsibility for the choices of artificial agents is shared between high-level and low-level agents.

My discussion has also shown that any resolution of the Moral Proxy Problem involves sacrifices: adopting the low-level perspective implies designers should make design decisions that have almost certainly worse aggregate outcomes than other available design decisions, and regulators should not step in to change this. Adopting the high-level perspective, on the other hand, involves designers or regulators imposing specific courses of action in matters where there is intuitively rational and moral leeway when human agents are involved and where, prior to the introduction of new technology, the state and tech companies exerted no such control. It also risks absolving users of artificial agents of felt or actual responsibility for the artificial agents they employ, and having them live with consequences of choices they would not have made.

Finally, I have shown that because the way in which expected utility theory is commonly understood and implemented in artificial agent design implies risk neutrality regarding goal satisfaction, it involves, in a sense, a tacit endorsement of the high-level response to the Moral Proxy Problem which makes such risk neutrality generally plausible. Given low-level agential framing, risk aversion is intuitively neither always irrational nor immoral, and is in fact common in human agents. The implication is that if we prefer a low-level response to the Moral Proxy Problem in at least some contexts, risk aversion should be made room for in the design of artificial agents. Whichever solution to the Moral Proxy Problem we settle on, I hope my discussion has at least shown that the largely unquestioned implementation of risk neutrality in the design of artificial agents deserves critical scrutiny and that such scrutiny reveals that the right treatment of risk is intimately connected with how we answer difficult questions about agential perspective and responsibility in a world increasingly populated by artificial agents.

5

Artificial Intelligence and Its Integration into the Human Lifeworld

*Christoph Durt**

I. INTRODUCTION

Artificial Intelligence (AI) is a rapidly advancing yet much misunderstood technology. Vastly different definitions of AI, ranging from AI as a mere tool to an intelligent being, give rise to contradicting assessments of the possibilities and dangers of AI. A clearer concept of AI is needed to come to a better understanding of the possibilities of responsible governance of AI. In particular, the relation of AI to the world we live in needs to be clarified. This chapter shows that AI integrates into the human lifeworld much more thoroughly than other technology, and that the integration needs to be understood within a wider picture.

The reasons for the unclear concept of AI do not merely lie in AI's novelty, but also in the fact that it is an extraordinary technology. This chapter will take a fresh look at the unique nature of AI. The concept of AI here is restricted to computational systems: hard- and software that make up devices and applications which may but do not usually resemble humans. This chapter rejects the common assumption that AI is necessarily a simulation or even replication of humans or of human capacities and explains that what distinguishes AI from other technologies is rather its special relation to the world we live in.

The world we live in includes ordinary physical nature, which humans have been extensively changing with the help of technology – in constructive and in destructive ways. Human life is constantly becoming more bound to technology, up to the degree that the consequences of the use of technology threaten the most fundamental conditions of life on earth. Even small conveniences provided by technology, such as taking a car or plane instead of a bicycle or public transportation, matter more to most of us than the environmental damage they cause. Our dependence on technology has become so self-evident that a standard answer to the problems caused by technology is that they will be taken care of by future technology.

Technology is not only changing the physical world, however, and this chapter elaborates why this is especially true for AI. The world we live in is also what philosophers since *Edmund*

* Work on this chapter was supported by the Volkswagen Foundation in the project 'AI and Its Integration into the World of Experience and Meaning' and the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 754340. Many of the thoughts in this chapter have been thoroughly discussed in the context of the former project, for which I would like to thank Christian Müller, Julian Nida-Rümelin, Philipp Slusallek, Maren Wehrle, and Nathalie Weidenfels. Special thanks to David Winkens for our in-depth discussions and his concrete feedback on this chapter. Further thanks to Daniel Feuerstack, Alisa Pojtinger, and Silja Voeneky for their helpful feedback on this chapter.

Husserl have called the 'lifeworld' (*Lebenswelt*).[1] As the 'world of actually experiencing intuition',[2] the lifeworld founds higher-level meaning-formation.[3] It is hence not only a 'forgotten meaning-fundament of natural science'[4] but the 'horizon of all meaningful induction'.[5] The lifeworld is not an assumed reality behind experience, but the world we actually experience, which is meaningful to us in everyday life. *Husserl* himself came from mathematics to philosophy, and the concept of lifeworld culminates his lifelong occupation with the relation between mathematics and experience. He elaborates in detail how, over the course of centuries, the lifeworld became 'mathematized'.[6] In today's expression, we may say that the lifeworld becomes 'digitized'.[7] This makes *Husserl's* concept of lifeworld especially interesting for AI. While *Husserl* was mostly concerned with the universal structures of experience, however, this chapter will use the concept of lifeworld in a wider sense that includes social and cultural structures of experience, common sense, and language, as well as rules and laws that order common everyday activities.

Much of technology becomes integrated into the lifeworld in the sense that its use becomes part of our ordinary lives, for example in the forms of tools we use. AI, however, also integrates into the lifeworld in an especially intimate way: by intelligently navigating and changing meaning and experience. This does not imply human-like intelligence, which involves consciousness and understanding. Rather, AI makes use of different means, which may or may not resemble human intelligence. What makes them intelligent is not their apparent resemblance to human capacities, but the fact that they navigate and change the lifeworld in ways that make sense to humans. For instance, a self-driving car must 'recognize' stop signs and act accordingly, but AI recognition may be very different from human recognition.

Conventional high-tech, such as nuclear power plants, does not navigate the space of human meaning and experience. Even technologies that aim at changing meaning and experience, such as TV and the Internet, will look primitive in comparison to future AI's active and fine-grained adaptation to the lifeworld. AI is set to disrupt the human lifeworld more profoundly than conventional technologies, not because it will develop consciousness and will, but because it integrates into the lifeworld in a way not known from previous technology. A coherent understanding of how AI technology relates to the world we live in is necessary to assess its possible uses, benefits, and dangers, as well as the possibilities for responsible governance of AI.

AI attends to and possibly changes not just physical aspects of the lifeworld, but also those of meaning and experience, and it does so in exceedingly elaborate, 'intelligent,' ways. Like other technology, AI takes part in many processes that do not directly affect the lifeworld. In contrast to other technology, however, AI integrates into the lifeworld in the just delineated special sense. Doing so had before been reserved to humans and animals. While it should be self-evident that AI does not need to use the same means, such as conscious understanding, the resemblance to human capacities has caused much confusion. It is probably the strongest reason for the typical conceptions

[1] E Husserl, *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy* (1970) (hereafter Husserl, *Crisis*).
[2] Ibid, 50.
[3] Ibid, 140 (*Geltungsfundierung*).
[4] Ibid, 48.
[5] Ibid, 50.
[6] For an extensive elaboration of the different steps involved in the 'mathematization of nature' and its relation to the lifeworld see C Durt, *The Paradox of the Primary-Secondary Quality Distinction and Husserl's Genealogy of the Mathematization of Nature* (2012).
[7] For a detailed elaboration, see C Durt, 'The Computation of Bodily, Embodied, and Virtual Reality' (2020) *Phänomenologische Forschungen* 25 (hereafter Durt, 'The Computation of Bodily, Embodied, and Virtual Reality').

of AI as a replication or simulation of human intelligence, conceptions that have misled the assessment of AI and lie behind one-sided enthusiasm and alarmism about AI. It is time to explore a new way of explaining how AI integrates into the lifeworld, as will be done in this chapter.

The investigation starts in Section II with an analysis of the two prevalent conceptions of AI in relation to the world. Traditionally and up to today, the relation of AI to the world is either thought to be that of an object in the world, such as a tool, or that of a subject that experiences and understands the world, or a strange mixture of object and subject. In particular, the concept of AI as a subject has attracted much attention. Already the Turing Test compares humans and machines as if they were different persons, and early visionaries believed AI could soon do everything a human can do. Today, a popular question is not whether AI will be able to simulate all intelligent human behaviour, but when it will be as intelligent as humans.

Section III argues that the subject and the object conception of AI both fundamentally misrepresent the relation of AI to the world. It will be shown that this has led to grave misconceptions of the chances and dangers of AI technology and has hindered both the development and assessment of AI. The attempt to directly compare AI with humans is deeply ingrained in the history of AI, and this chapter analyses in detail how the direct comparison plays out already in the setup of the Turing Test.

Section IV shows that the Turing Test allows for intricate exchanges and is much harder on the machine than it appears at first sight. By making the evaluator part of the experiment, the Turing Test passes on the burden of evaluation, but does not remove it.

The multiple roles of the evaluator are differentiated in Section V. Making the evaluator part of the test covers up the difference between syntactic data and the semantic meaning of data, and it hides in plain sight that the evaluator adds the understanding that is often attributed to the AI. We need a radical shift of perspective that looks beyond the core computation of the AI and considers how the AI itself is embedded in the wider system, which includes the lifeworld.

Section VI will map out further the novel approach to the relation of AI to lifeworld. It elaborates how humans and AI relate to the lifeworld in very different ways. The section explores how the interrelations of AI with humans and data enable AI to represent and simulate the lifeworld. In their interaction, these four parts constitute a whole that allows a better understanding of the place of AI.

II. THE OBJECT AND THE SUBJECT CONCEPTION OF AI

While in today's discussions of AI there is a widespread sense that AI will fundamentally change the world we live in, assessments of the growing impact of AI on the world differ widely. The fundamental disagreements already start with the definition of AI. There is a high degree of uncertainty about whether AI is a technology comparable to objects such as tools or machines, or to subjects of experience and understanding, such as humans.

Like other technologies, AI is often reduced to material objects, such as tools, devices, and machines, which are often simply called 'technology', together with the software that runs on them. The technological processes in which material technological devices take part, however, are also called 'technology.' This latter use is closer to the Greek root of technology, *technē* (*τέχνη*), which refers to particular kinds of making or doing. Today, 'technology' is primarily used to refer to technological hard- and software and only secondarily to their use. To refer to the hard- and software that makes up an AI, this chapter will simply speak of an 'AI'. The application of conventional concepts to AI makes it look as if there were only two fundamentally different possibilities to conceive of the relation of AI to the world: that of (1) an object and (2) a subject.

The first takes AI to be an object such as a tool or a machine and assesses its impact on the world we live in in the same categories as that of conventional technologies. It is certainly true that AI can be part of devices we can use for certain purposes, good or bad. Because tools enable and suggest certain uses, and disable or discourage others, they are not neutral objects. Tools are objects that are embedded in a use, which means that they mediate the relationship of humans to the world.[8] These are important aspects of AI technology. The use of technology cannot be ignored and there are attempts to focus on the interaction between material objects and their use, such as 'material engagement theory.'[9] The chapter at hand affirms that such theories take a step in the right direction and yet shows that they do not go far enough to understand the nature of AI. It is not wrong to say that AI systems are material objects that are used in certain ways (if 'material object' includes data and software), but this does not suffice to account for this novel technology. While conceiving AI as a mere object used in particular ways is true in some respects, it does not tell the whole story.

AI exhibits features we do not know from any conventional technology. Devices that make use of AI can do things otherwise only known from the intelligent and autonomous behaviour of humans and sometimes animals. AI systems can process large amounts of meaningful data and use it to navigate the lifeworld in meaningful ways. They can perform functions so complex they are hard to fathom but need to be explained in ordinary language.[10] AI systems are not mere objects in the world, nor are they only objects that are used in particular ways, such as tools. Rather, they actively relate to the world in ways that often would require consciousness and understanding if humans were to do them.[11] AI here changes subjective aspects of the lifeworld, although it does not necessarily experience or understand, or simulate experience or understanding. The object concept of AI ignores the fact that AI can operate on meaningful aspects of the world and transform them in meaningful ways. No other technology in the history of humanity has done so. AI indeed entails enormous potential – both to do good and to inflict harm.

The subject concept of AI (2) attempts to account for the fact that AI can do things we otherwise only know from humans and animals. AI is imagined as a being that relates to the world analogously to a living subject: by subjectively experiencing and understanding the world by means of mental attitudes such as beliefs and desires. The most common form of the subject account of AI is the idea that AI is something more or less like a human, and that it will possibly develop into a super-human being. Anthropomorphic conceptions of AI are often based on an animistic view of AI, according to which software has a mind, together with the materialistic view that brains are computers.[12] Some proponents who hold this view continue the science-fiction narrative of aliens coming to earth.[13] Vocal authors claim that AI will at one point in time be intelligent in the sense that it will develop a mind of its own. They think that Artificial General Intelligence (AGI) will engage in high-level mental activities and claim that computers will literally attain consciousness and develop their own will. Some speculate that this may

[8] D Ihde, *Technology and the Lifeworld: From Garden to Earth* (1990).

[9] D Ihde and L Malafouris, 'Homo Faber Revisited: Postphenomenology and Material Engagement Theory' (2019) 32 *Philosophy & Technology* 195.

[10] The need to explain complex AI processing to non-experts has given rise to a whole new field of study, that of 'Explainable AI.'

[11] MO Riedl, 'Human-Centered Artificial Intelligence and Machine Learning' (*Arxiv*, 31 January 2019) http://arxiv.org/abs/1901.11184.

[12] J Nida-Rümelin and N Weidenfeld, *Digitaler Humanismus: Eine Ethik für das Zeitalter der künstlichen Intelligenz* (2018).

[13] N Weidenfeld, 'Wo Bleibt Der Mensch?' (2019) 3 *Neue Gesellschaft Frankfurter Hefte* 16.

happen very soon, in 2045,[14] or at least well before the end of this century.[15] Estimates like these are used to justify either enthusiastic salvation phantasies,[16] or alarmistic warnings of the end of humanity.[17] In the excitement caused by such speculations, however, it is often overlooked that they promote a concept of AI that has more to do with science fiction than actual AI science.

Speculative science fiction phantasies are only one, extreme, expression of the subject conception of AI. The next section investigates the origin of the subject conception of AI in the claim that AI can simulate human intelligence. The comparison with natural intelligence is already suggested by the term AI, and the next sections investigate why the comparison of human and artificial intelligence has misled thinking on AI. There is a sense in which conceiving of AI as a subject is due to a lack rather than a hypertrophy of phantasy: the lack of imagination when it comes to alternative ways of understanding the relation of AI to the world. The basic problem with the object and subject conceptions of AI is that they apply old ways of thinking to a novel technology that calls into question old categories such as that of object and subject. Because these categories are deeply rooted in human thought, they are hard to overcome. In the next section, I argue that the attempt to directly compare them is misleading and the resulting confusion prone to hinder both the development and assessment of AI.

III. WHY THE COMPARISON OF HUMAN AND ARTIFICIAL INTELLIGENCE IS MISLEADING

Early AI researchers did not try to artificially recreate consciousness but rather to simulate human capabilities. Today's literal ascriptions of behaviour, thinking, experience, understanding, or authorship to machines ignore a distinction that was already made by the founders of the study of 'Artificial Intelligence.' AI researchers such as *John McCarthy*, *Marvin Minsky*, *Nathaniel Rochester*, and *Claude Shannon* were aware of the difference between, on the one hand, thinking, experiencing, understanding and, on the other, their simulation.[18] They did not claim that a machine could be made that could literally understand. Rather, they thought that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[19]

If AI merely simulates some human capacities, it is an object with quite special capacities. For many human capacities, such as calculating, this prospect is relatively unexciting and does not contradict the object idea of AI. The idea that AI can simulate core or even all features of intelligence, however, gives the impression of some mixture of subject and object, an uncanny 'subject-object.'[20] In the case of *McCarthy et al.*,[21] the belief in the powers of machine intelligence goes along with a belief that human intelligence is reducible to the workings of a machine. But this is a very strong assumption that can be questioned in many respects. Is it

[14] R Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (2005), (hereafter Kurzweil, 'The Singularity').
[15] According to the predominant view of 170 persons who responded to a request to 549 'experts' selected by VC Müller and N Bostrom, 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion' in VC Müller (ed), *Fundamental Issues of Artificial Intelligence* (2016).
[16] Kurzweil, 'The Singularity' (n 14).
[17] N Bostrom, 'Existential Risk Prevention as Global Priority' (2013) 4 *Global Policy* 15.
[18] J McCarthy and others, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence' (1955) https://rockfound.rockarch.org/digital-library-listing/-/asset_publisher/yYxpQfeI4W8N/content/proposal-for-the-dartmouth-summer-research-project-on-artificial-intelligence (hereafter J McCarthy and others 1955)
[19] Ibid, 1.
[20] C Durt, 'The Computation of Bodily, Embodied, and Virtual Reality'.
[21] J McCarthy and others 1955 (n 18).

really true that all aspects of learning and all other features of intelligence can be precisely described? Are learning and intelligence free of ambiguity and vagueness?

Still today, nearly 70 years later, and despite persistent efforts to precisely describe learning and intelligence, there is no coherent computational account of all learning and intelligence. Newer accounts often question the assumption that learning and intelligence are reduceable to computation.[22] While more caution with regard to bold prophecies would seem advisable, the believers in AGI do not think that the fact that we have no general computational account of intelligence speaks against AGI. They believe that, even though we do not know how human intelligence works, we will somehow be able to artificially recreate general intelligence. Or, if not us, then the machines themselves will do it. This is not only deemed a possibility, but a necessity. The fact that such beliefs are based on speculation and not a clear concept of their alleged possibility is hidden behind seemingly scientific numerical calculations that produce precise results such as the number '2045' for the year in which the supposedly certain and predictable event of 'singularity' will happen, which is when the development of 'superhuman intelligence' becomes uncontrollable and leads to the end of the 'human era'.[23]

In the 1950s and 1960s, the idea that AI should simulate human intelligence was not far off from the efforts of actual AI research. Today, however, most real existing AI does not even attempt to simulate human behaviour and thinking. Only a small part of AI research attempts to give the appearance of human intelligence, although that part is still disproportionally represented in the media. For the most widely used AI technologies, such as Machine Learning (ML), this is not the case. The reason is obvious: machines are most effective not when they attempt to simulate human behaviour but when they make full use of their own strengths, such as the capability to process vast amounts of data in little time. The idea that AI must simulate human intelligence has little to do with the actual development of AI. Even more disconnected from reality are the speculations around the future rise of AGI and its potential consequences.

Yet, even in serious AI research, such as on ML, the tendency to think of AI in comparison to humans persists. When ML is covered, then often by using comparisons to human intelligence that are easily misleading, such as "system X is better than humans in recognizing Y." Such claims tend to conceal that there are very specific conditions under which the ML system is better than humans. 'Recognition' is defined with respect to input-output relations. The machine is made the measure of all things. It is conveniently overlooked that current ML capabilities break down already in apparently straightforward 'recognition' tasks when there are slight changes to the input. The reason is simple: the ML system is usually not doing the same thing humans do when they recognize something. Rather, it uses means such as data correlation to replace recognition tasks or other work that had before been done by humans – or to accomplish things that before had not been possible or economic. Clearly, none of this means that ML becomes human-like. Even in social robotics, it is not always conducive for social interaction to build robots that resemble humans as closely as possible. One disadvantage is expressed in the concept of 'uncanny valley' (or 'uncanny cliff'[24]), which refers to the foundering

[22] C Tewes, C Durt, and T Fuchs, 'Introduction: The Interplay of Embodiment, Enaction, and Culture' in C Durt, T Fuchs, and C Tewes (eds), *Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World* (2017).

[23] V Vinge, 'The Coming Technological Singularity: How to Survive in the Post-Human Era' (1993) https://ntrs.nasa.gov/api/citations/19940022856/downloads/19940022856.pdf.

[24] C Bartneck and others, 'Is the Uncanny Valley an Uncanny Cliff?' (The 16th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2007, Korea, September 2007).

of acceptance of humanoid robots when their close resemblance evokes eerie feelings. Claiming that AI systems are becoming human-like makes for sensationalistic news but does not foster clear thought on AI.

While AI systems are sometimes claimed to be better than humans at certain tasks, they have obvious troubles when it comes to 'meaning, reasoning, and common-sense knowledge',[25] all of which are fundamental to human intelligence. On the other hand, ML in particular can process inhuman amounts of data in little time. If comparisons of AI systems with humans makes sense at all, then only with reservations and with regard to aspects of limited capabilities. Because of the vast differences between the capabilities, AI is not accurately comparable to a human, not even to an x-year-old child.

For the above reasons, the definitions of AI as something that simulates or replicates human intelligence are misleading. Such anthropomorphic concepts of AI are not apt to understand and assess AI. We need a radically different approach that better accounts for how AI takes part in the world we live in. A clear understanding of the unique ways in which AI is directed to the lifeworld does not only allow for a better assessment of the impact of AI on the lifeworld but is furthermore crucial for AI research itself. AI research suffers from simplistic comparisons of artificial and human intelligence, which make progress seem alternatively very close or unreachable. Periods in which it seems as if AI would soon be able to do anything a human can do alternate with disappointment and the drying out of funding ('AI Winter'[26]). Overcoming of the anthropomorphic concept of AI contributes to more steady progress in AI science.

How natural it is for humans to reduce complex developments to simplistic notions of agency is obvious in animistic conceptions of natural events and in conspiracy theories. Because AI systems show characteristics that appear like human agency, perception, thought, or autonomy, it is particularly tempting to frame AI in these seemingly well-known terms. This is not necessarily a problem as long as it is clear that AI cannot be understood in analogy to humans. Exactly this is frequently suggested, however, by the comparison of AI systems to humans with respect to some capacity such as 'perception.' Leaving behind the idea that AI needs to be seen in comparison to natural intelligence allows us to consider anew how different AI technologies such as ML can change, disrupt, and transform processes by integrating into the lifeworld. But this is easier said than done. The next section shows how deeply rooted the direct comparison of humans and AI is in the standard account of AI going back to *Alan Turing*.

IV. THE MULTIPLE ROLES OF THE EVALUATOR IN THE TURING TEST

The Turing Test is the best-known attempt to conceive of a quasi-experimental setting to find out whether a machine is intelligent or not.[27] Despite its age – *Turing* published the thought experiment, later called the 'Turing Test', in 1950 – it is still widely discussed today. It can serve as an illustrative example for the direct comparison of AI to humans and how this overlooks their specific relations to the world. In this section I argue that by making the 'interrogator' part of the experiment, the Turing Test only seemingly avoids difficult philosophical questions.

Figure 5.1 shows a simple diagram of the Turing Test. A human and the AI machinery are put in separate rooms and, via distinct channels, exchange text messages with an 'interrogator' who

[25] S Lohr, 'Is There a Smarter Path to Artificial Intelligence? Some Experts Hope So' (2018) *The New York Times*, www.nytimes.com/2018/06/20/technology/deep-learning-artificial-intelligence.html.

[26] D Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence* (1993) 203.

[27] AM Turing, 'Computing Machinery and Intelligence' (1950) 59 *Mind* 433 (hereafter Turing, 'Computing Machinery').

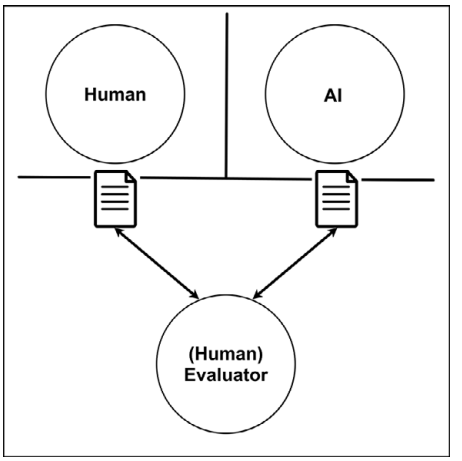

FIGURE 5.1 The Turing Test

has, apart from the content of the messages, no clues as to which texts stem from the human and which from the machine. The machine is built in such a way that the answers it gives to the questions of the 'interrogator' appear as if they were from a human. It competes with the human in the other room, who is supposed to convince the evaluator that their exchange is between humans. The 'interrogator' is not only asking questions but is furthermore tasked to judge which of the entities behind the respective texts is a human. If the human 'interrogator' cannot correctly distinguish between the human and the machine, the machine has passed the Turing Test.

The Turing Test is designed to reveal, in a straightforward way, whether a machine can be said to exhibit intelligence. By limiting the exchanges to texts, the design puts the human and the AI on the same level. This allows for a direct comparison of their respective outputs. As pointed out by *Turing* himself, the direct comparison with humans may be unfair to the machine because it excludes the possibility that the machine develops other kinds of intelligence that may not be recognized by the human 'interrogator'.[28] It furthermore does not take into consideration potential intelligent capabilities of the AI that do not express in exchanges of written text. On the side of the human, the restriction to text exchanges excludes human intelligence that cannot be measured in text exchanges. Textual exchanges are just one of many forms in which intelligent behaviour and interaction of humans may express itself. While the limitation to textual exchanges enables a somewhat 'fair' evaluation, at the same time it distorts the comparison.

At first sight, the Turing Test seems to offer only a few possibilities of interaction by means of texts. *Turing's* description of the test suggests that the 'interrogator' merely ask questions and the human or the AI give answers. Already interrogation can consist of extreme vetting and involve a profound psychological examination as well as probing of the consistence of the story unveiled in the interrogation.[29] Furthermore, there is nothing in the setup that limits the possible interchanges to questions and answers. The text exchanges may go back and forth in myriad ways. The 'interrogator' is as well a conversation partner who engages in the text-driven

[28] Turing, 'Computing Machinery' (n 27).
[29] J Landgrebe and B Smith, 'There Is No Artificial General Intelligence' (*Arxiv*, 9 June 2019) http://arxiv.org/abs/1906.05833.

conversations. He or she takes on, at the same time, the multiple roles of interrogator, reader, interpreter, interlocuter, conversation partner, evaluator, and judge. This chapter uses the wider term 'evaluator', which is less restrictive than 'interrogator' and is meant to comprise all mentioned roles.

Like other open-ended text exchanges, the Turing Test can develop in intricate ways. *Turing* was surely aware of the possible intricacies of the exchanges because the declared origin of his test is the 'Imitation Game.'[30] The Imitation Game involves pretending to be of a different gender, a topic *Turing* may have been confronted with in his own biography. If we conceive the exchanges in terms of Wittgenstein's concept of language-games, it is clear that the rules of the language game are usually not rigid but malleable and sometimes can be changed in the course of the language-game itself.[31] In free exchanges that involve 'creative rule-following',[32] the interchange may seem to develop on its own due to the interplay of possibly changing motivations, interests, and emotions, as well as numerous natural and cultural factors. While the intricate course of a conversation often seems logical in hindsight, it can be hard to predict even for humans, and exceedingly so for those who do not share the same background and form of life.

Mere prediction of probable words can result in texts that make sense up to a certain degree. Without human editing, they may appear intelligent in the way a person can appear intelligent who rambles on about any trigger word provided by the 'conversation' 'partner'. It is likely to leave the impression of somebody or something that did not understand or listen to the other. Text prediction is not sufficient to engage in a genuine conversation. The claim that today's advanced AI prediction systems such as GPT-3 are close to passing the Turing Test[33] are much exaggerated as long as the test is not overly limited by external factors such as a narrow time frame, or a lack of intelligence, understanding, and judgement on the part of the human evaluator.

The Turing Test is thus as much a test of the 'intelligence' of an AI system as it is a test of how easy (or hard) it is to trick a human into believing that some machine-generated output constitutes a text written by a human. That was probably the very idea behind the Turing test: tricking a human into believing that one is a human is a capability that surely requires intelligence. The fact that outside of the Turing Test it is often astonishingly easy to trick a human into believing there was an intelligent being behind some action calls into question the idea that humans always show an impressive 'intelligence.' The limitations of human intelligence can hence make it easier for a machine to pass a Turing Test. The machine could also simply attempt to pretend to be a human with limited language capabilities. On the other hand, however, faking human flaws can be very difficult for machines. Human mistakes and characteristics such as emotional reactions or tiredness are natural to humans but not to machines and may prove difficult to simulate.[34] If the human evaluator is empathetic, he or she is likely to have a feeling for emotional states expressed in the texts. Thus, not only the intellectual capabilities of the evaluator but also their, in today's expression, 'emotional intelligence' plays a role. All of this may seem self-evident for humans, which is why it may be easy to overlook how much the Turing Test asks of the evaluator.

[30] A Turing, 'Computing Machinery,' 433.
[31] L Wittgenstein, *Philosophische Untersuchungen: Kritisch-genetische Edition* (4th 2001) §83.
[32] C Durt, 'From Calculus to Language Game: The Challenge of Cognitive Technology' (2018) 22 *Techné: Research in Philosophy and Technology* 425 (hereafter Durt, 'From Calculus to Language Game').
[33] T Taulli, 'Turing Test at 70: Still Relevant for AI (Artificial Intelligence)?' (*Forbes*, 27 November 2020) www.forbes.com/sites/tomtaulli/2020/11/27/turing-test-at-70-still-relevant-for-ai-artificial-intelligence/.
[34] Durt, 'From Calculus to Language Game' (n 32).

Considering the intricate exchanges possible in the Turing Test, the simplicity of its setup is deceptive. *Turing* set up the test in a way that circumvents complicated conceptual issues involved in the question 'can machines think?' It only does so, however, because it puts the burden of evaluation on the evaluator, who needs to figure out whether the respective texts are due to intelligence or not. If, however, we attempt to unravel the exact relations between the evaluator, the other human, the machine, the texts, and the world, we are back to the complicated conceptual, philosophical, and psychological questions *Turing* attempted to circumvent with his test.

The evaluator may not know that such questions are implicitly involved in her or his evaluation and instead may find the decision obvious or decide by gut feeling. But the better the machine simulates a human and the more difficult it becomes to distinguish it from a human, the more relevant for the evaluation becomes a differentiated consideration of the conditions of intelligence. Putting the burden of decision on the evaluator or anybody else does not solve the complicated conceptual issues that are brought up by machines that appear intelligent. For the evaluator, the process of decision is only in so far simplified that the setup of the Turing Test prevents her or him from inspecting the outward appearance or the internal workings of the machine. The setup frames the evaluation, which also means that it may mislead the evaluation by hiding in plain sight the contribution of the evaluator.

V. HIDDEN IN PLAIN SIGHT: THE CONTRIBUTION OF THE EVALUATOR

While the setup of the Turing Test puts the burden of assessing whether a machine is intelligent on the evaluator, it also withholds important information from the evaluator. Because it prevents the evaluator from knowing anything about the processes behind the outputs, one can always imagine that some output was produced by means other than understanding. We need to distinguish two meanings of 'intelligent' and avoid the assumption that the one leads to the other. 'Intelligent' in the first sense concerns the action, which involves understanding of the meaning of the task. Task-solving without understanding the task, for example, by looking up the solutions in the teacher's manual, is usually not called an 'intelligent' solution of the task, at least not in the same sense.

The other sense of 'intelligent' refers to the solution itself. In this sense, the solution can be intelligent even when it was produced by non-intelligent means. Because the result is the same as that achieved by understanding, and the evaluator in the Turing Test only gets to see the results, he or she is prevented from distinguishing between the two kinds of intelligence. At the same time, however, the design suggests that intelligence in the second sense amounts to intelligence in the first sense. The Turing Test replaces the question 'Can machines think?' with the 'closely related'[35] question whether a machine can react to input with output that makes a human believe it thinks. In effect, Turing demands that if the output is deemed intelligent (in the second sense), then the machine should be called intelligent (in the first sense). Due to the setup of the Turing Test, this can only be a pragmatic criterion and not a proof. It is no wonder the Turing Test has led to persistent confusions. The confusion of the two kinds of 'intelligent' and confusions with regard to the interpretation of the Turing Test are pre-programmed in its setup.

Especially confusing is the source of the meaning of the texts the evaluator receives. On the one hand, the texts may appear to be produced in an understanding manner, on the other hand,

[35] Turing, 'Computing Machinery' (n 27) 433.

the evaluator is withheld any knowledge of how they were produced. In general, to understand texts, their constituting words and symbols must not only be recognized as such but also be understood.[36] In the Turing Test, it is the human evaluator who reads the texts and understands their meaning. Assumedly, the human in one room, too, understands what the texts mean, but the setup renders irrelevant whether this really is the case. Both the human and the machine may not have understood the texts they produced. The only thing that matters is whether the evaluator believes that the respective texts were produced by a human. The evaluator will only believe that the texts were produced by a human, of course, when they appear to express an understanding of their semantic meaning. The fact that the texts written by the human and produced by the machine need to be interpreted is easily overlooked because the interpretation is an implicit part of the setup of the Turing Test. By interpreting the texts, the evaluator adds the meaning that is often ascribed to the AI output.

The texts exchanged in the Turing Test have very different relevance for humans and for computers. For digital computation, the texts are relevant only with respect to their syntax. They constitute mere sets of data, and data only in its syntactic form, regardless of what it refers to in the world, or, indeed, whether it refers to anything. For humans, data means more than syntax. Like information, data is a concept that is used in fundamentally different ways. Elsewhere I distinguished different senses of information,[37] but for reasons of simplicity this chapter speaks only of data, and only two fundamentally different concepts of data will be distinguished.

On the one hand, the concept of data is often used syntactically to signify symbols stored at specific memory locations that can be computationally processed. On the other hand, the concept of data is used semantically to signify meaningful information about something. Semantic meaning of data paradigmatically refers to the world we live in, such as the datum '8,849' for the approximate height of Mount Everest. Data can represent things, relations, and temporal developments in the world, including human bodies, and they may also be used to simulate aspects of the real or a potentially existing world. Furthermore, data can represent language that is not limited to representative data. Humans only sometimes use data to engage in a communication and talk about something in the shared world. They are even less often concerned with the syntactic structure of data. Quite often, texts can convey all kinds of semantic content: besides information, they can convey moods, inspire fantasy, cause insight, produce feelings, and challenge the prejudices of their readers.

To highlight that data can be used to represent complex structures of all kinds, I here also speak of 'digital knowledge.' Like data, there is a syntactic and a semantic meaning of digital knowledge. Computers operate on syntactic relations of what constitutes semantic knowledge once it appropriately represents the world. The computer receives syntactic data as an input and then processes the data according to syntactic rules to deliver a syntactically structured output. Syntactic data processing can be done in different ways, for example, by means of logical gates, neuronal layers, or quantum computing. Despite the important differences between these methods of data processing, they are still syntactic methods of data processing, of course. Data processing is at the core of computational AI.

If we want to consider whether a computer has intelligence by itself, the fundamental question is whether certain transformations of data can constitute intelligence. Making the machine look like a human does not fundamentally change the question. In this regard, Turing is justified when he claims that there 'was little point in trying to make a "thinking machine"

[36] Cf. D Carr, *The Paradox of Subjectivity: The Self in the Transcendental Tradition* (1999), 124.
[37] Ibid.

more human by dressing it up in such artificial flesh.'[38] Most likely, doing so would only lead to complications and confusions. Because the core work of computing is syntactic symbol-manipulation, the restriction to texts is appropriate with regard to the core workings of computational AI.

The intelligence to be found here, however, can only concern the second sense of 'intelligent' that does not involve semantic understanding. The fact that mere syntactic operations are not sufficient for semantic understanding has been pointed out by numerous philosophers in the context of different arguments. *Gottfried Wilhelm Leibniz* holds 'that perception and that which depends upon it are inexplicable on mechanical grounds.'[39] *John Searle* claims that computers 'have syntax but no semantics,'[40] which is the source of the 'symbol grounding problem.'[41] *Hubert Dreyfus* contends that there are certain things a certain kind of AI, such as 'symbolic AI,' can never do.[42] Recent researchers contend that 'form [. . .] cannot in principle lead to learning of meaning.'[43] All these arguments do not show that there is no way to syntactically model understanding, but rather that no amount of syntactic symbol-manipulation by itself amounts to semantic understanding. There is no semantic understanding in computation alone. The search for semantic understanding in the computational core of AI looks at the wrong place.

The point of this chapter is not to contribute another argument for the negative claim that there is something computation cannot do. The fundamental difference between syntactic data and semantic meaning does not mean that syntactic data cannot map structures of semantic meaning or that it could not be used to simulate understanding behaviour. According to *Husserl*, data can 'dress' the lifeworld like a 'garb of ideas,'[44] which fits the lifeworld so well that the garb is easily mistaken for reality in itself. Because humans are also part of reality, it easily seems like the same must be possible for humans. Vice versa, data can cause behaviour (e.g., of robots) that sometimes resembles human behaviour in such a perfect manner that it looks like conscious behaviour. At least with regard to certain behaviours it is possible that an AI will appear like a human in the Turing Test or even in reality, even though this is much harder than usually thought.[45]

The point of differentiating between syntactic computation and semantic meaning in this chapter is to build bridges rather than to dig trenches. To understand how the two can cooperate, we need to understand how they are embedded in a wider context. Although it is futile to look for meaning and understanding in the computational core of AI, this is not the end of the story. Even when AI systems by themselves do not experience and understand, they may take part in a wider context that also comprises other parts. To make progress on the question of how AI can meaningfully integrate into the lifeworld, it is crucial to shift the perspective away from the computational AI devices and applications alone toward the AI in its wider context.

[38] Turing, 'Computing Machinery' (n 27) 434.

[39] GW Leibniz and N Rescher, *G.W. Leibniz's Monadology: An Edition for Students* (1992) 83.

[40] JR Searle, 'Minds, Brains, and Programs' (1980) 3 *Behavioral and Brain Sciences* 417, 423; cf. JR Searle, 'The Problem of Consciousness' (1993) 2 *Consciousness and Cognition* 310, where Searle holds that computers do not even have syntax.

[41] S Harnad, 'The Symbol Grounding Problem' (1990) 42 *Physica D: Nonlinear Phenomena* 335.

[42] HL Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (1992); cf. SE Dreyfus and HL Dreyfus, 'Towards a Reconciliation of Phenomenology and AI' in D Partridge and Y Wilks (eds), *The Foundations of Artificial Intelligence: A Sourcebook* (1990) for a more optimistic view.

[43] EM Bender and A Koller, 'Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data' in *Association for Computational Linguistics* (eds), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020) 5185.

[44] E Husserl, *Crisis*, 51.

[45] See Section IV.

The Turing Test can again serve as an example of the embeddedness of the AI in a wider context. By withholding from the evaluator any knowledge of how the texts are processed, the Turing Test stands in the then-prevailing tradition of behaviourism. The Turing Test sets up a 'black box' in so far as it hides from the evaluator all potentially relevant information and interaction apart from what is conveyed in the texts. By making the evaluator part of the test, however, Turing goes beyond classical behaviourism. The content of the texts may enable inferences to the mental processes of the author such as motivations and reasoning processes, inferences which the evaluator is likely to use to decide whether there is a human behind the respective channel. By allowing such inferences, the Turing Test is closer to cognitivism than behaviourism. Yet, making the evaluator part of the setup is not a pure form of cognitivism either. To come to a decision, the (human) evaluator needs to understand the meaning of the texts and reasonably evaluate them. By making the evaluator part of the test, understanding of semantic meaning becomes an implicit part of the test. The setup of the Turing Test as a whole constitutes a bigger system, of which the AI is only one part. The point here is not that the system as a whole would understand or be intelligent, but that only because the texts are embedded in the wider system, they are meaningful texts rather than mere objects.

Data is another important part of that bigger system. For the AI in the Turing Test, the input and output texts constitute syntactic data, whereas for the evaluator they have semantic meaning. The semantic meaning of data goes beyond language and refers to things in the world we live in. The lifeworld is hence another core part of the bigger system and needs to be considered in more detail.

VI. THE OVERLOOKED LIFEWORLD

The direct comparison of AI with humans overlooks the fact that AI and humans relate to the lifeworld in very different ways, which is the topic of this section. As mentioned in the introduction, AI systems such as autonomous cars need not only navigate the physical world but also the lifeworld. They need to recognize a stop sign as well as the intentions of other road users such as pedestrians who want to cross the road, and act or react accordingly. In more abstract terms, they need to be able to recognize and use the rules they encounter in their environment, together with regulations, expectations, demands, logic, laws, dispositions, interconnections, and so on.

Turing had recognized that the development and use of intelligence is dependent on things that shape how humans are embedded in, and conceive of, the world, such as culture, community, emotion, and education.[46] Nevertheless, and despite the incompatibilities discussed in the last section, behaviourists and cognitivists assume that in the Turing Test all of these can be ignored when probing whether a machine is intelligent or not. They overlook that the texts exchanged often refer to the world, and that their meaning needs to be understood in the context of what they say about the world. Because the texts consist outwardly only of data, they need to be interpreted by somebody to mean something.[47] By interpreting the texts to mean something, the evaluator adds meaning to the texts, which would otherwise be mere collections of letters and symbols. Here, the embeddedness of the evaluator into the lifeworld – including culture, community, emotion, and education – as well as inferences to the lifeworld of the human behind the channel come into play.

[46] AM Turing, 'Intelligent Machinery' in BJ Copeland (ed), *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma* (2004) 430–431.

[47] See Section V.

The realization that the limitation to textual exchanges captures only part of human intelligence has led to alternative test suggestions. *Steve Wozniak*, the co-founder of Apple Inc., proposed the AI should enter the house of a stranger and make a cup of coffee.[48] The coffee test is an improvement over the Turing Test in certain respects. The setup of the coffee test does not hide the relation of the output of the AI to the lifeworld; to the contrary, it explicitly chooses a task that seems to require orientation in the lifeworld. It involves an activity that is relatively easy for humans (but may be quite intricate), who can make use of their common-sense knowledge and reason, to find their way in a stranger's home. While the Turing Test may involve common sense, for instance to understand or to answer certain questions, this involvement is not as obvious as it is in the coffee test. There are several questions about the coffee test, however. The action of making coffee is much simpler than engaging in open-ended exchanges of meaningful text and may be solved in ways that in fact do only require limited orientation in the physical world rather than general orientation in the lifeworld. Most important in our context is, however, that, like the Turing Test, *Wozniak's* test still attempts to directly compare AI with human capabilities. As argued above, this is not apt to adequately capture the strengths of AI and is likely to lead to misrepresentations of the relation of the AI system to the lifeworld.

The relation of the AI system to the lifeworld is mediated through input and output consisting of data, regardless of whether the data corresponds to written texts or is provided by and transmitted to interfaces. Putting a robotic body around a computational system does make a difference in that it enables the system to retrieve data from sensors and interfaces in relation to movements initiated by computational processing. But it doesn't change the fact that, like any computational system, ultimately the robot continues to relate to the lifeworld by means of data. The robotic sensors provide it with data, and data is used to steer the body of the robot, but data alone is not sufficient for experience and understanding. Like any machine, the robot is a part of the world, but it does not have the same intentional relation to the world. Humans literally experience the lifeworld and understand meaning, whereas computational AI does not literally do so – not even when the AI is put into a humanoid robot. The outward appearance that the robot relates to the world like a living being is misleading. Computational AI thus can never be integrated in the lifeworld in the same way humans are. Yet, it would plainly be wrong to claim that they do not relate to the lifeworld at all.

Figure 5.2 shows the fundamental relations between humans and AI together with their respective relations to data and the lifeworld that are described in this chapter. Humans literally experience the lifeworld and understand meaning, whereas computational AI receives physical sensor input from the lifeworld and may modify physical aspects of the lifeworld by means of connected physical devices. AI can (1) represent and (2) simulate the lifeworld, by computing (syntactic) data and digital knowledge that corresponds to things and relations in the lifeworld. The dotted lines indicate that data and digital knowledge do not represent and simulate by themselves. Rather, they do so by virtue of being appropriately embedded in the overall system delineated in Figure 5.2. The AI either receives sensor or interface input that is stored in digital data and which can be computationally processed and used to produce output. The output can be used to modify aspects of the lifeworld, for example, to control motors or interfaces that are accessible to other computing systems or to humans.

[48] M Shick, 'Wozniak: Could a Computer Make a Cup of Coffee?' (*Fast Company*, 2 March 2010) www.fastcompany.com/1568187/wozniak-could-computer-make-cup-coffee; see also B Goertzel, M Iklé, and J Wigmore, 'The Architecture of Human-Like General Intelligence' in P Wang and B Goertzel (eds), *Theoretical Foundations of Artificial General Intelligence*, vol 4 (2012).

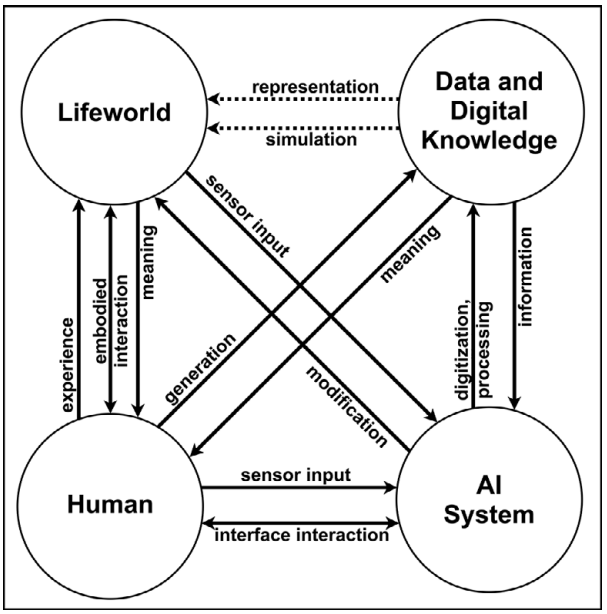

FIGURE 5.2 The fundamental relations between humans, AI, data, and the lifeworld

In contrast to mere computation, which only operates with data, AI in addition needs to intelligently interact in the lifeworld. Even in the Turing Test, in spite of the restriction of the interactions to textual interchange rather than embodied interactions, the lifeworld plays a crucial role. From the discussion of the test, we can extract two main reasons: (1) the textual output needs to make sense in the context of the lifeworld of the evaluator, and (2) even though exchanges by means of written texts are rather limited, the exchange is carried out via a channel in the lifeworld of the evaluator. The interchange itself happens in the lifeworld, and the AI needs to give the impression of engaging in the interchange.

In other applications of AI technology, AI devices are made to intelligently navigate and modify the lifeworld, as well as interact in it. As pointed out in Section VI, autonomous cars need to take into account the behaviour of human road users such as human drivers and pedestrians. The possibly relevant behaviour of human road users is generally neither the result of strict rule-following nor is it just random. Humans often behave according to their understanding of the situation, their aims, perspectives, experiences, habits, conventions, etc. Through these, humans direct themselves to the lifeworld. Humans experience the lifeworld as meaningful, and their behaviour (mostly) makes sense to other members of the same culture.

Relating to the lifeworld in intelligent ways is an exceptionally difficult undertaking for computational AI because it needs to do so by means of data processing. As acknowledged above, there is a radical difference between syntactic data processing and experience and understanding, a difference that cannot be eliminated by more syntactic data processing. This radical difference comes through in the difference of, on the one hand, data and digital knowledge, and, on the other, the lifeworld. As discussed above, syntactic data and digital knowledge need to be interpreted to say something about the lifeworld, but such interpretation cannot be done by data processing. The difference between syntactic data processing and experience and understanding must be bridged, however, if AI is supposed to intelligently interact in the lifeworld. The combination of the need to bridge the difference and the

impossibility of doing so by data processing alone looks like an impasse if we limit the view to the AI and the processing of data and digital knowledge.

We are not stuck in the apparent impasse, however, if we take into account the wider system. The wider system bridges the gap between, on the one hand, data and digital knowledge, and, on the other, the lifeworld. Bridging a gap is different from eliminating a gap, as the difference between both sides remains. All bridges are reminders of the obstacle bridged. Bridges are pragmatic solutions that involve compromises and impose restrictions. In our case, data and digital knowledge do not fully capture the world as it is experienced and understood but rather represent or simulate it. The representation and simulation are made possible by the interplay of the four parts delineated in the diagram.

Biomimetics can certainly inspire new engineering solutions in numerous fields, and AI research especially is well-advised to take a more careful look at how human cognition really operates. I argued above that a naïve understanding of human cognition has led to misguided assessments of the possibilities of AI. The current section has given reasons for why a better understanding of human cognition needs to take into account how humans relate to the lifeworld.

It would be futile, however, to exactly rebuild human cognition by computational means. As argued in Section III, the comparison of human and artificial intelligence has led to profound misconceptions about AI, such as those discussed in Sections IV and V. The relation of humans to their lifeworld matters for AI research, not because AI can fully replace humans but because AI relates to the lifeworld in particular ways. To better understand how AI can meaningfully integrate into the lifeworld, the role of data and digital knowledge needs to be taken into account, and the interrelations need to be distinguished in the way delineated in Figure 5.2. This is the precondition for a prudent assessment of both the possibilities and dangers of AI and to envision responsible uses of AI in which technology and humans do not work against but with each other.

PART II

Current and Future Approaches to AI Governance

6

Artificial Intelligence and the Past, Present, and Future of Democracy

*Mathias Risse**

I. INTRODUCTION: HOW AI IS POLITICAL

Langdon Winner's classic essay 'Do Artifacts Have Politics?' resists a widespread but naïve view of the role of technology in human life: that technology is neutral, and all depends on use.[1] He does so without enlisting an overbearing determinism that makes technology the sole engine of change. Instead, *Winner* distinguishes two ways for artefacts to have 'political qualities'. First, devices or systems might be means for establishing patterns of power or authority, but the design is flexible: such patterns can turn out one way or another. An example is traffic infrastructure, which can assist many people but also keep parts of the population in subordination, say, if they cannot reach suitable workplaces. Secondly, devices or systems are strongly, perhaps unavoidably, tied to certain patterns of power. *Winner's* example is atomic energy, which requires industrial, scientific, and military elites to provide and protect energy sources. Artificial Intelligence (AI), I argue, is political the way traffic infrastructure is: It can greatly strengthen democracy, but only with the right efforts. Understanding 'the politics of AI' is crucial since *Xi Jinping's* China loudly champions one-party rule as a better fit for our digital century. AI is a key component in the contest between authoritarian and democratic rule.

Unlike conventional programs, AI algorithms learn by themselves. Programmers provide data, which a set of methods, known as machine learning, analyze for trends and inferences. Owing to their sophistication and sweeping applications, these technologies are poised to dramatically alter our world. Specialized AI is already broadly deployed. At the high end, one may think of AI mastering Chess or Go. More commonly we encounter it in smartphones (Siri, Google Translate, curated newsfeeds), home devices (Alexa, Google Home, Nest), personalized customer services, or GPS systems. Specialized AI is used by law enforcement, the military, in browser searching, advertising and entertainment (e.g., recommender systems), medical diagnostics, logistics, finance (from assessing credit to flagging transactions), in speech recognition producing transcripts, trade bots using market data for predictions, but also in music creations and article drafting (e.g., GPT-3's text generator writing posts or code). Governments track people using AI in facial, voice, or gait recognition. Smart cities analyze traffic data in real time or design services. COVID-19 accelerated use of AI in drug discovery. Natural language

* I am grateful to audiences at University College London and at the University of Freiburg for helpful discussions during Zoom presentations of this material in June 2021. I also acknowledge helpful comments from Sushma Raman, Derya Honca, and Silja Voeneky.

[1] L Winner, 'Do Artifacts Have Politics?' (1980) 109 *Daedalus* 121.

processing – normally used for texts – interprets genetic changes in viruses. Amazon Web Services, Azure, or Google Cloud's low- and no-code offerings could soon let people create AI applications as easily as websites.[2]

General AI approximates human performance across many domains. Once there is general AI smarter than we are, it could produce something smarter than itself, and so on, perhaps very fast. That moment is the singularity, an intelligence explosion with possibly grave consequences. We are nowhere near anything like that. Imitating how mundane human tasks combine agility, reflection, and interaction has proven challenging. However, 'nowhere near' means 'in terms of engineering capacities'. A few breakthroughs might accelerate things enormously. Inspired by how millions of years of evolution have created the brain, neural nets have been deployed in astounding ways in machine learning. Such research indicates to many observers that general AI will emerge eventually.[3]

This essay is located at the intersection of political philosophy, philosophy of technology, and political history. My purpose is to reflect on medium and long-term prospects and challenges for democracy from AI, emphasizing how critical a stage this is. Social theorist *Bruno Latour*, a key figure in Science, Technology and Society Studies, has long insisted no entity matters in isolation but attains meaning through numerous, changeable relations. Human activities tend to depend not only on more people than the protagonists who stand out, but also on non-human entities. *Latour* calls such multitudes of relations actor-networks.[4] This perspective takes the materiality of human affairs more seriously than is customary, the ways they critically involve artefacts, devices, or systems. This standpoint helps gauge AI's impact on democracy.

Political theorists treat democracy as an ideal or institutional framework, instead of considering its materiality. Modern democracies involve structures for collective choice that periodically empower relatively few people to steer the social direction for everybody. As in all forms of governance, technology shapes how this unfolds. Technology explains how citizens obtain information that delineates their participation (often limited to voting) and frees up people's time to engage in collective affairs to begin with. Devices and mechanisms permeate campaigning and voting. Technology shapes how politicians communicate and bureaucrats administer decisions. Specialized AI changes the materiality of democracy, not just in the sense that independently given actors deploy new tools. AI changes how collective decision making unfolds and what its human participants are like: how they see themselves in relation to their environment, what relationships they have and how those are designed, and generally what forms of human life can come to exist.[5]

[2] For current trends, see P Chojecki, *Artificial Intelligence Business: How You Can Profit from AI* (2020). For the state of the art, see M Mitchell, *Artificial Intelligence: A Guide for Thinking Humans* (2019); T Taulli, *Artificial Intelligence Basics: A Non-Technical Introduction* (2019); S Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (2019). See also The Future Today Institute, '14th Annual Tech Trends Report' (2021); for musings on the future of AI, see J Brockman (ed), *Possible Minds: Twenty-Five Ways of Looking at AI* (2019).

[3] For optimism about the occurrence of a singularity, see R Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (2006); for pessimism, see EJ Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (2021); see also N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2016) (hereafter Bostrom, *Superintelligence*); M Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (2017) (hereafter Tegmark, *Life 3.0*).

[4] B Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory* (2007); B Latour, *We Have Never Been Modern* (1993). To be sure, and notwithstanding the name of the theory, *Latour* speaks of actants rather than actors, to emphasize the role of non-human entities.

[5] How to understand 'technology' is a non-trivial question in the philosophy of technology, as it affects how broad our focus is; see C Mitcham, *Thinking through Technology: The Path between Engineering and Philosophy* (1994); M Coeckelbergh, *Introduction to Philosophy of Technology* (2019). For AI one could just think of a set of tools in

Section II explores what democracy is, emphasizes the materiality of 'early' and 'modern' democracy and rearticulates the perspective we take from Winner. Section III recounts some of the grand techno-skeptical narratives of twentieth-century philosophy of technology, distilling the warnings they convey for the impact of AI on democracy. Section IV introduces another grand narrative, a Grand Democratic AI Utopia, a way of imagining the future we should be wary of. Section V discusses challenges and promises of AI for democracy in this digital century without grand narratives. Instead, we ask how to design AI to harness the public sphere, political power, and economic power for democratic purposes, to make them akin to *Winner's* inclusive traffic infrastructure. Section VI concludes.

II. DEMOCRACY AND TECHNOLOGY

A distinctive feature – and an intrinsic rather than comparative advantage – of recognizably democratic structures is that they give each participant at least minimal ownership of social endeavors and inspire many of them to recognize each other as responsible agents across domains of life. There is disagreement about that ideal, with *Schumpeterian* democracy stressing peaceful removal of rulers and more participatory or deliberative approaches capturing thicker notions of empowerment.[6] Arguments for democracy highlight democracy's possibilities for emancipation, its indispensability for human rights protection, and its promise of unleashing human potentials. Concerns to be overcome include shortsightedness vis-a-vis long-term crises, the twin dangers of manipulability by elites and susceptibility to populists, the potential of competition to generate polarization, and a focus on process rather than results. However, a social-scientific perspective on democracy by *David Stasavage* makes it easier to focus on its materiality and thus, later on, the impact of AI.[7] *Stasavage* distinguishes early from modern democracy, and both of those from autocracy. Autocracy is governance without consent of those people who are not directly controlled by the ruling circles anyway. The more viable and thus enduring autocracies have tended to make up for that lack of consent by developing a strong bureaucracy that would at least guarantee robust and consistent governance patterns.

1. *Early Democracy and the Materiality of Small-Scale Collective Choice*

Early democracy was a system in which rulers governed jointly with councils or assemblies consisting of members who were independent from rulers and not subject to their whims. Sometimes such councils and assemblies would provide information, sometimes they would assist with governance directly. Sometimes councils and assemblies involved participation from large parts of the population (either directly or through delegation), sometimes councils were

machine learning; alternatively, one could think of the whole set of devices in which these tools are implemented, and all productive activities that come with procurement and extraction of materials involved; see K Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (2021) (hereafter Crawford, *Atlas of AI*). While I mostly sideline these issues, I adopt an understanding of technology from W Bijker, 'Why and How Technology Matters' in RE Goodin and C Tilly (eds), *The Oxford Handbook of Contextual Political Analysis* (2006). At a basic level, 'technology' refers to sets of *artefacts* like computers, cars, or voting machines. At the next level, it also includes human *activities*, as in 'the technology of e-voting'. Thereby it refers also to the making and handling of such machines. Finally, and closest to its Greek origin, 'technology' refers to *knowledge*: It is about what people know as well as what they do with machines and related production processes.

[6] For a good overview, see A Gutmann, 'Democracy' in RE Goodin, P Pettit, and TW Pogge (eds), *A Companion to Contemporary Political Philosophy* (2007).

[7] D Stasavage, *The Decline and Rise of Democracy: A Global History from Antiquity to Today* (2020) (hereafter Stasavage, *The Decline and Rise of Democracy*).

elite gatherings. Rulership might be elective or inherited. Its material conditions were such that early democracy would arise in smaller rather than larger polities, in polities where rulers depended on subjects for information about what they owned or produced and so could not tax without compliance, and where people had exit options. Under such conditions, rulers needed consent from parts of the population. Early democracy thus understood was common around the globe and not restricted to Greece, as the standard narrative has it.[8]

However, what is special about Athens and other Greek democracies is that they were most extensively participatory. The reforms of *Cleisthenes*, in the sixth century BC, divided Athens into 139 demes (150 to 250 men each, women playing no political role) that formed ten artificial 'tribes'. Demes in the same tribe did not inhabit the same region of Attica. Each tribe sent 50 men, randomly selected, for a year, to a Council of 500 to administer day-to-day affairs and prepare sessions of the Assembly of all citizens. This system fed knowledge and insights from all eligible males into collective decision making without positioning anyone for take-over.[9] It depended on production and defense patterns that enslaved people to enable parts of the population to attend to collective affairs. Transport and communication had to function to let citizens do their parts. This system also depended on a steady, high-volume circulation of people in and out of office to make governance impersonal, representative, and transparent at the same time. That flow required close bookkeeping to guarantee people were at the right place – which involved technical devices, the material ingredients of democratic governance.

Let me mention some of those. The kleroterion (allotment machine) was a two-by-three-foot slab of rock with a grid of deep, thin slots gouged into it. Integrating some additional pieces, this sophisticated device helped select the required number of men from each tribe for the Council, or for juries and committees where representation mattered. Officers carried allotment tokens – pieces of ceramics inscribed with pertinent information that fit with another piece at a secure location to be produced if credentials were questioned. (Athens was too large for everyone to be acquainted.) With speaking times limited, a water clock (*klepsydra*) kept time. Announcement boards recorded decisions or messages. For voting, juries used ballots, flat bronze disks. Occasionally, the Assembly considered expelling citizens whose prominence threatened the impersonal character of governance, ostracisms for which citizens carved names into potsherds. *Aristotle* argued that citizens assembled for deliberation could display virtue and wisdom no individual could muster, an argument for democracy resonant through the ages.[10] It took certain material objects to make it work. These objects were at the heart of Athenian democracy, devices in actor-networks to operationalize consent of the governed.[11]

[8] To think of Greek democracy as a uniquely located innovation also contradicts the evolutionary story of early bands of humans who succeeded because they were good at cooperating and had brains that had evolved to serve cooperative purposes. See for example, C Boehm, *Hierarchy in the Forest. The Evolution of Egalitarian Behavior* (1999). To the extent that a demos separate from an aristocracy is the hallmark of democracy (a sensible view given the etymology), many cases covered by *Stasavage* do not count. Still, his account creates an illuminating contrast with autocracies. Also, in structures where consent is needed, internal dynamics over time typically demand broader inclusion.

[9] Stasavage, *The Decline and Rise of Democracy* (n 7) 29; J Ober, *The Rise and Fall of Classical Greece* (2015) 123; J Thorley, *Athenian Democracy* (2004) 23.

[10] O Höffe (ed), *Aristotle. Politics* (1998). Also see M Risse, 'The Virtuous Group: Foundations for the 'Argument from the Wisdom of the Multitude'' (2001) 31 *Canadian Journal of Philosophy* 31, 53.

[11] For the devices, I draw on J Dibbell, 'Info Tech of Ancient Democracy' (*Alamut*), www.alamut.com/subj/artiface/deadMedia/agoraMuseum.html, which explores museum literature on these artefacts displayed in Athens. See also S Dow, 'Aristotle, the Kleroteria, and the Courts' (1939) 50 *Harvard Studies in Classical Philology* 1. For the mechanics of Athenian democracy, see also MH Hansen, *The Athenian Democracy in the Age of Demosthenes: Structure, Principles, and Ideology* (1991).

2. Modern Democracy and the Materiality of Large-Scale Collective Choice

As a European invention, modern democracy is representative, with mandates that do not bind representatives to an electorate's will. Representatives emerge from competitive elections under increasingly universal suffrage. Participation is broad but typically episodic. The material conditions for its existence resemble early democracy: they emerge where rulers need subjects to volunteer information and people have exit options. But modern democracies arise in large territories, as exemplified by the United States.[12] Their territorial dimensions (and large populations) generate two legitimacy problems. First, modern democracy generates distrust because 'state' and 'society' easily remain abstract and distant. Secondly, there is the problem of overbearing executive power. Modern democracies require bureaucracies to manage day-to-day-affairs. Bureaucracies might generate their own dynamics, and eventually citizens no longer see themselves governing. If the head of the executive is elected directly, excessive executive power becomes personal power.[13]

Modern democracy too depends on material features to function. Consider the United States. In 1787 and 1788, *Alexander Hamilton*, *James Madison*, and *John Jay*, under the collective pseudonym 'Publius', published 85 articles and essays ('Federalist Papers') to promote the constitution. *Hamilton* calls the government the country's 'center of information'.[14] 'Information' and 'communication' matter greatly to Publius: the former term appears in nineteen essays, the latter in a dozen. For these advocates of this trailblazing system, the challenge is to find structures for disclosure and processing of pertinent information about the country. Publius thought members of Congress would bring information to the capital, after aggregating it in the states. But at the dawn of the Republic, the vastness of the territory made these challenges formidable. One historian described the communication situation as a 'quarantine' of government from society.[15] Improvements in postal services and changes in the newspaper business in the nineteenth century brought relief, facilitating the central role of media in modern democracies. Only such developments could turn modern democracies into actor-networks where representatives do not labor in de-facto isolation.[16]

'The aim of every political constitution is or ought to be first for rulers to obtain men who possess most wisdom to discern, and most virtue to pursue the common good of society', we read in Federalist No. 57.[17] To make this happen, in addition to a political culture where the right people seek office, voting systems are required, the design of which was left to states. Typically,

[12] *Hélène Landemore* has argued that modern democracy erred in focusing on representation. Instead, possibilities of small-scale decision making with appropriate connections to government should have been favored – which now is more doable through technology. See H Landemore, 'Open Democracy and Digital Technologies' in L Bernholz, H Landmore, and R Reich (eds), *Digital Technology and Democratic Theory* (2021) 62; H Landemore, *Open Democracy: Reinventing Popular Rule for the Twenty-First Century* (2020).

[13] *Howard Zinn* has a rather negative take specifically on the founding of the United States that would make it unsurprising that these legitimacy problems arose: 'Around 1776, certain important people in the English colonies [. . .] found that by creating a nation, a symbol, a legal unity called the United States, they could take over land, profits, and political power from favorites of the British Empire. In the process, they could hold back a number of potential rebellions and create a consensus of popular support for the rule of a new, privileged leadership'; H Zinn, *A People's History of the United States* (2015) 59.

[14] JE Cooke, *The Federalist* (1961), 149 (hereafter Cooke, *Federalists*).

[15] JS Young, *The Washington Community 1800–1828* (1966) 32.

[16] B Bimber, *Information and American Democracy: Technology in the Evolution of Political Power* (2003) 89. For the argument that, later, postal services were critical to the colonization of the American West (and thus have been thoroughly political throughout their existence), see C Blevins, *Paper Trails: The US Post and the Making of the American West* (2021).

[17] Cooke, *Federalist* (n 14) 384.

what they devised barely resembled the orderliness of assigning people by means of the *kleroterion*. 'Ballot' comes from Italian *ballotta* (little ball), and ballots often were something small and round, like pebbles, peas, beans or bullets.[18] Paper ballots gradually spread, partly because they were easier to count than beans. Initially, voters had to bring paper and write down properly spelled names and offices. The rise of parties was facilitated by that of paper ballots. Party leaders printed ballots, often in newspapers – long strips, listing entire slates, or pages to be cut into pieces, one per candidate. Party symbols on ballots meant voters did not need to know how to write or read, an issue unknown when people voted by surrendering beans or by voice.

In 1856, the Australian state of Victoria passed its Electoral Act, detailing the conduct of elections. Officials had to print ballots and erect booths or hire rooms. Voters marked ballots secretly and nobody else was allowed in polling places. The 'Australian ballot' gradually spread, against much resistance. Officially, such resistance arose because it eliminated the public character of the vote that many considered essential to honorable conduct. But de facto there often was resistance because the Australian ballot made it hard for politicians to get people to vote for them in exchange for money (as such voting behavior then became hard to verify). In 1888, Massachusetts passed the first statewide Australian-ballot law in the United States. By 1896, most Americans cast secret, government-printed ballots. Such ballots also meant voters had to read, making voting harder for immigrants, formerly enslaved people, and the uneducated poor. Machines for casting and counting votes date to the 1880s. Machines could fail, or be manipulated, and the mechanics of American elections have remained contested ever since.

3. *Democracy and Technology: Natural Allies?*

The distant-state and overbearing-executive problems are so substantial that, for *Stasavage*, 'modern democracy is an ongoing experiment, and in many ways, we should be surprised that it has worked at all.'[19] The alternative to democracy is autocracy, which is viable only if backed by competent bureaucracies. *Stasavage* argues that often advances in production and communication undermined early democracy. New or improved technologies could reduce information advantages of subjects over rulers, e.g., regarding fertility of land – if governments have ways of assessing the value of land, they know to tax it; if they do not, they have no good way of taxing it without informational input from the owners. Agricultural improvements led to people living closer together so bureaucrats could easily monitor them. Conversely, slow progress in science and development favored survival of early democracy.

Innovations in writing, mapping, measurement, or agriculture made bureaucracies more effective, and thus made autocracies with functioning bureaucracies the more viable. Much depends on sequencing. Entrenched democracies are less likely to be undermined by technological advances than polities where autocracy is a live option. And so, in principle, entrenched democracies these days could make good use of AI to enhance their functionality (and thus make AI a key part of the materiality of contemporary democracies). In China, the democratic alternative

[18] I follow J Lepore, 'Rock, Paper, Scissors: How We Used to Vote' (*The New Yorker*, 13 October 2008). Some of those themes also appear in J Lepore, *These Truths: A History of the United States* (2019), especially chapter 9. See also RG Saltman, *History and Politics of Voting Technology: In Quest of Integrity and Public Confidence*. (2006). For the right to vote in the United States, see A Keyssar, *The Right to Vote: The Contested History of Democracy in the United States* (2009).

[19] Stasavage, *The Decline and Rise of Democracy* (n 7) 296. For a political-theory idealization of modern democracy in terms of two 'tracks', see J Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (1996) Chapters 7–8. The first track is formal decision making (e.g., parliament, courts, agencies). The other is informal public deliberation, where public opinion is formed.

never gained much traction. In recent decades, the country made enormous strides under an autocratic system with a competent bureaucracy. Under *Xi Jinping*, China aggressively advertises its system, and AI has started to play a major role in it, especially in the surveillance of its citizens.[20]

Yuval Noah Harari recently offered a somewhat different view of the relationship between democracy and technology.[21] Historically, he argues, autocracies have faced handicaps around innovation and growth. In the late twentieth century especially, democracies outperformed dictatorships because they were better at processing information. Echoing *Hayek's Road to Serfdom*, *Harari* thinks twentieth-century technology made it inefficient to concentrate information and power.[22] But *Harari* also insists that, at this stage, AI might altogether alter the relative efficiency of democracy vs. authoritarianism.

Stasavage and *Harari* agree that AI undermines conditions that make democracy the more viable system. This does not mean existing democracies are in imminent danger. In fact, it can only be via technology that individuals matter to politics in modern democracies in ways that solve the distant-state and overbearing-executive problems. Only through the right kind of deployment of modern democracy's materiality could consent to governance be meaningful and ensure that governance in democracies does not mean quarantining leadership from population, as it did in the early days of the American Republic. As the twenty-first century progresses, AI could play a role in this process. Because history has repeatedly shown how technology strengthens autocracy, democrats must be vigilant vis-à-vis autocratic tendencies from within. Technology is indispensable to make modern democracy work, but it is not its natural ally. Much as in *Winner's* infrastructure design, careful attention must be paid to ensure technology advances democratic purposes.[23]

III. DEMOCRACY, AI, AND THE GRAND NARRATIVES OF TECHNO-SKEPTICISM

Several grand techno-skeptical narratives have played a significant role in the twentieth-century philosophy of technology. To be sure, that field now focuses on a smaller scale, partly because grand narratives are difficult to establish.[24] However, these narratives issue warnings about how

[20] The success of the Chinese model has prompted some philosophers to defend features of that model, also in light of how democracies have suffered from the two legitimacy problems; see DA Bell, *The China Model: Political Meritocracy and the Limits of Democracy* (2016); T Bai, *Against Political Equality: The Confucian Case* (2019); J Chan, *Confucian Perfectionism: A Political Philosophy for Modern Times* (2015). For the view that China's Communist Party will face a crisis that will force it to let China become democratic, see Ci, *Democracy in China: The Coming Crisis* (2019). For the argument that different governance models emerge for good reasons at different times, see F Fukuyama, *The Origins of Political Order: From Pre-Human Times to the French Revolution* (2012); F Fukuyama, *Political Order and Political Decay: From the Industrial Revolution to the Globalization of Democracy* (2014).

[21] YN Harari, 'Why Technology Favors Tyranny' (*The Atlantic*, October 2018) www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.

[22] FA Hayek, *The Road to Serfdom* (2007).

[23] Similarly, Cohen and Fung – reviewing deterministic viewpoints that see technology clearly favor or disfavor democracy – conclude that 'the democratic exploitation of technological affordances is vastly more contingent, more difficult, and more dependent on ethical conviction, political engagement, and good design choices than the technological determinists appreciated' A Fung and J Cohen, 'Democracy and the Digital Public Sphere' in L Bernholz, H Landemore, and R Reich (eds), *Digital Technology and Democratic Theory* (2021) 25 (hereafter Fung and Cohen, 'Democracy and the Digital Public Sphere'). Or as computer scientist Nigel Shadbolt says, addressing worries that 'machines might take over': '[T]he problem is not that machines might wrest control of our lives from the elites. The problem is that most of us might never be able to wrest control of the machines from the people who occupy the command posts', N Shadbolt and R Hampson, *The Digital Ape: How to Live (in Peace) with Smart Machines* (2019) 63.

[24] M Coeckelbergh, *Introduction to Philosophy of Technology* (2019) Part II.

difficult it might be to integrate specifically AI into flourishing democracies, warnings we are well advised to heed as much is at stake.

1. *Lewis Mumford and the Megamachine*

Mumford was a leading critic of the machine age.[25] His 1934 *Technics and Civilization* traces a veritable cult of the machine through Western history that often devastated creativity and independence of mind.[26] He argues that 'men had become mechanical before they perfected complicated machines to express their new bent and interest'.[27] People had lived in coordinated ways (forming societal machines) and endorsed ideals of specialization, automation, and rationality before physical machines emerged. That they lived that way made people ready for physical machines. In the Middle Ages, mechanical clocks (whose relevance for changing life patterns *Mumford* tirelessly emphasizes) literally synchronized behavior.[28]

Decades later *Mumford* revisited these themes in his two-volume 'Myth of the Machine.[29] These works offer an even more sweeping narrative, characterizing modern doctrines of progress as scientifically upgraded justifications for practices the powerful had deployed since pharaonic times to maintain power. Ancient Egypt did machine work without actual machines.[30] Redeploying his organizational understanding of machines, *Mumford* argues pyramids were built by machines – societal machines, centralized and subtly coordinated labor systems in which ideas like interchangeability of parts, centralization of knowledge, and regimentation of work are vital. The deified king, the pharaoh, is the chief engineer of this original megamachine. Today, the essence of industrialization is not even the large-scale use of machinery. It is the domination of technical knowledge by expert elites, and our structured way of organizing life. By the early twentieth century, the components of the contemporary megamachine were assembled, controlled by new classes of decision makers governing the 'megatechnical wasteland' (a dearth of creative thinking and possibilities in designing their own lives on the part of most people).[31] The 'myth' to be overcome is that this machine is irresistible but also beneficial to whoever complies.

Mumford stubbornly maintained faith in life's rejuvenating capacities, even under the shadow of the megamachine. But clearly any kind of AI, and social organization in anticipation of general AI, harbors the dangers of streamlining the capacities of most people in society that *Mumford* saw at work since the dawn of civilization. This cannot bode well for governance based on meaningful consent.

2. *Martin Heidegger and the World As Gestell*

Heidegger's most influential publication on technology is his 1953 *'The Question Concerning Technology'*.[32] Modern technology is the contemporary mode of understanding things.

[25] On Mumford, see DL Miller, *Lewis Mumford: A Life* (1989).
[26] L Mumford, *Technics and Civilization* (2010).
[27] Ibid,, 3.
[28] Ibid, 12–18.
[29] L Mumford, *Myth of the Machine: Technics and Human Development* (1967) (hereafter Mumford, *Myth of the Machine*); L Mumford, *Pentagon of Power: The Myth of the Machine* (1974) (hereafter Mumford, *Pentagon of Power*).
[30] Mumford, *Myth of the Machine* (n 29) chapter 9.
[31] The title of chapter 11 of Mumford, *Pentagon of Power* (n 29).
[32] M Heidegger, *The Question Concerning Technology, and Other Essays* (1977) 3–35 (hereafter Heidegger, *The Question Concerning Technology*). On Heidegger, see J Richardson, *Heidegger* (2012); ME Zimmerman, *Heidegger's Confrontation with Modernity: Technology, Politics, and Art* (1990).

Technology makes things show up as mattering, one way or another. The mode of revealing (as *Heidegger* says) characteristic of modern technology sees everything around us as merely a standing-reserve (*Bestand*), resources to be exploited as means.[33] This includes the whole natural world, even humans. In 1966, *Heidegger* even predicted that 'someday factories will be built for the artificial breeding of human material'.[34]

Heidegger offers the example of a hydroelectric plant converting the Rhine into a mere supplier of waterpower.[35] In contrast, a wooden bridge that has spanned the river for centuries reveals it as a natural environment and permits natural phenomena to appear as objects of wonder. *Heidegger* uses the term *Gestell* (enframing) to capture the relevance of technology in our lives.[36] The prefix 'Ge' is about linking together of elements, like *Gebirge*, mountain range. Gestell is a linking together of things that are posited. The Gestell is a horizon of disclosure according to which everything registers only as a resource. Gestell deprives us of any ability to stand in caring relations to things. Strikingly, *Heidegger* points out that 'the earth now reveals itself as a coal mining district, the soil as a material deposit'.[37] Elsewhere he says the modern world reveals itself as a 'gigantic petrol station'.[38] Technology lets us relate to the world only in impoverished ways. Everything is interconnected and exchangeable, efficiency and optimization set the stage. Efficiency demands standardization and repetition. Technology saves us from having to develop skills while also turning us into people who are satisfied with lives that do not involve many skills.

For *Heidegger*, modern democracy with its materiality could only serve to administer the Gestell, and thus is part of the inauthentic life it imposes. His interpreter *Hubert Dreyfus* has shown how specifically the Internet exemplifies *Heidegger's* concerns about technology as Gestell.[39] As AI progresses, it would increasingly encapsulate *Heidegger's* worries about how human possibilities vanish through the ways technology reveals things. Democracies that manage to integrate AI should be wary of such loss.

3. *Herbert Marcuse and the Power of Entertainment Technology*

Twentieth-century left-wing social thought needed to address why the revolution as *Marx* predicted it never occurred. A typical answer was that capitalism persevered by not merely dominating culture, but by deploying technology to develop a pervasive entertainment sector. The working class got mired in consumption habits that annihilated political instincts. But Marxist thought sustains the prospect that, if the right path were found, a revolution would occur. In the 1930s, *Walter Benjamin* thought the emerging movie industry could help unite the masses in struggle, capitalism's efforts at cultural domination notwithstanding. Shared movie experiences could allow people to engage the vast capitalist apparatus that intrudes upon their daily lives. Deployed the right way, this new type of art could help finish up capitalism after all.[40]

When *Marcuse* published his 'One-Dimensional Man' in 1964, such optimism about the entertainment sector had vanished. While he had not abandoned the Marxist commitment to

[33] Heidegger, *The Question Concerning Technology* (n 32) 17.
[34] Quoted in Young, *Heidegger's Later Philosophy* (2001) 46.
[35] Heidegger, *The Question Concerning Technology* (n 32) 16.
[36] Ibid, 19.
[37] Ibid, 14.
[38] Quoted in J Young, *Heidegger's Later Philosophy* (2001) 50.
[39] HL Dreyfus, *On the Internet* (2008).
[40] W Benjamin, *The Work of Art in the Age of Its Technological Reproducibility, and Other Writings on Media* (2008).

the possibility of a revolution, *Marcuse* saw culture as authoritarian. Together, capitalism, technology, and entertainment culture created new forms of social control, false needs and a false consciousness around consumption. Their combined force locks one-dimensional man into one-dimensional society, which produces the need for people to recognize themselves in commodities. Powers of critical reflection decline. The working class can no longer operate as a subversive force capable of revolutionary change.

'A comfortable, smooth, reasonable, democratic unfreedom prevails in advanced civilization, a token of technical progress', *Marcuse* starts off.[41] Technology – as used especially in entertainment, at which *Benjamin* still looked differently – immediately enters *Marcuse's* reckoning with capitalism. It is 'by virtue of the way it has organized its technological base, [that] contemporary industrial society tends to be totalitarian'.[42] He elaborates: 'The people recognize themselves in their commodities; they find their soul in their automobile, hi-fi set, split-level home, kitchen equipment.'[43] Today, *Marcuse* would bemoan that people see themselves in possibilities offered by AI.

4. *Jacques Ellul and Technological Determinism*

Ellul diagnoses a systemic technological tyranny over humanity. His most celebrated work on philosophy of technology is 'The Technological Society'.[44] In the world *Ellul* describes, individuals factor into overall tendencies he calls 'massification'. We might govern particular technologies and exercise agency by operating machines, building roads, or printing magazines. Nonetheless, technology overall – as a *Durkheimian* social fact that goes beyond any number of specific happenings – outgrows human control. Even as we govern techniques (a term *Ellul* uses broadly, almost synonymously with a rational, systematic approach, with physical machines being the paradigmatic products), they increasingly shape our activities. We adapt to their demands and structures. *Ellul* is famous for his thesis of the autonomy of technique, its being a closed system, 'a reality in itself [. . .] with its special laws and its own determinations.' Technique elicits and conditions social, political, and economic change. It is the prime mover of all the rest, in spite of any appearances to the contrary and in spite of human pride, which pretends that man's philosophical theories are still determining influences and man's political regimes decisive factors in technical evolution.[45]

For example, industry and military began to adopt automated technology. One might think this process resulted from economic or political decisions. But for *Ellul* the sheer technical possibility provided all required impetus for going this way. *Ellul* is a technological determinist, but only for the modern age: technology, one way or another, causes all other aspects of society and culture. It does not have to be this way, and in the past it was not. But now, that is how it is.

Eventually, the state is inextricably intertwined with advancements of technique, as well as with corporations that produce machinery. The state no longer represents citizens if their interests contradict those advancements. Democracy fails, *Ellul* insists: we face a division between technicians, experts, and bureaucrats, standard bearers of techniques, on the one hand, and politicians who are supposed to represent the people and be accountable on the other.

[41] H Marcuse, *One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society* (1991) 1 (hereafter Marcuse, *One-Dimensional Man*).

[42] Marcuse, *One-Dimensional Man* (n 41) 3.

[43] Ibid, 9.

[44] J Ellul, *The Technological Society* (1964) (hereafter Ellul, *The Technological Society*). For recent discussions, see JP Greenman, *Understanding Jacques Ellul* (2012); HM Jerónimo, JL Garcia, and C Mitcham, *Jacques Ellul and the Technological Society in the 21st Century* (2013).

[45] Ellul, *The Technological Society* (n 44) 133.

'When the technician has completed his task,' *Ellul* says, 'he indicates to the politicians the possible solutions and the probable consequences – and retires.'[46] The technical class understands technique but is unaccountable. In his most chilling metaphor, *Ellul* concludes the world technique creates is 'the universal concentration camp'.[47] AI would perfect this trend.

IV. THE GRAND DEMOCRATIC AI UTOPIA

Let us stay with grand narratives a bit longer and consider what we might call the Grand Democratic AI Utopia. We are nowhere near deploying anything like what I am about to describe. But once general AI is on our radar, AI-enriched developments of *Aristotle's* argument from the wisdom of the multitude should also be. Futurists *Noah Yuval Harari* and *Jamie Susskind* touch on something like this;[48] and with technological innovation, our willingness to integrate technology into imageries for the future will only increase. Environmentalist *James Lovelock* thinks cyborgs could guide efforts to deal with climate change.[49] And in his discussion of future risks, philosopher *Toby Ord* considers AI assisting with our existential problems.[50] Such thinking is appealing because our brains evolved for the limited circumstances of small bands in earlier stages of *homo sapiens* rather than the twenty-first century's complex and globally interconnected world. Our brains could create that world but might not be able to manage its existential threats, including those we created.

But one might envisage something like this. AI knows everyone's preferences and views and provides people with pertinent information to make them competent participants in governance. AI connects citizens to debate views; it connects like-minded people but also those that dissent from each other. In the latter case, people are made to hear each other. AI gathers the votes, which eliminates challenges of people reaching polling stations, vote counting, etc. Monitoring everything, AI instantly identifies fraud or corruption, and flags or removes biased reporting or misleading arguments. AI improves procedural legitimacy through greater participation while the caliber of decision making increases because voters are well-informed. Voters no longer merely choose one candidate from a list. They are consulted on multifarious issues, in ways that keep them abreast of relevant complexities, ensure their views remain consistent, etc. More sophisticated aggregation methods are used than simple majoritarian voting.[51]

Perhaps elected politicians are still needed for some purposes. But by and large AI catapults early democracy into the twenty-first century while solving the problems of the distant state and of overbearing executive power. AI resolves relatively unimportant matters itself, consulting representative groups for others to ensure everything gets attention without swallowing too much time. In some countries citizens can opt out. Others require participation, with penalties for those with privacy settings that prohibit integration into the system. Nudging techniques – to get people to do what is supposed to be in their own best interest – are perfected for smooth operations.[52] AI avoids previously prevalent issues around lack of inclusiveness. Privacy settings protect all data. AI calls for elections if confidence in the government falls below a threshold.

[46] Ellul, *The Technological Society* (n 44) 258.
[47] Ibid, (n 44) 397.
[48] J Susskind, *Future Politics: Living Together in a World Transformed by Tech* (2018) Chapter 13; YN Harari, *Homo Deus: A Brief History of Tomorrow* (2018) Chapter 9.
[49] J Lovelock, *Novacene: The Coming Age of Hyperintelligence* (2020).
[50] See T Ord, *The Precipice: Existential Risk and the Future of Humanity* (2021) Chapter 5.
[51] For a discussion of majority rule in the context of competing methods that process information differently, also M Risse, 'Arguing for Majority Rule' (2004) 12 *Journal of Political Philosophy* 41.
[52] RH Thaler and CR Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (2009).

Bureaucracies are much smaller because AI delivers public services, evaluating experiences from smart cities to create smart countries. Judges are replaced by sophisticated algorithms delivering even-handed rulings. These systems can be arranged such that many concerns about functionality and place of AI in human affairs are resolved internally. In such ways enormous amounts of time are freed up for people to design their lives meaningfully.

As a desirable possibility, something like this might become more prominent in debates about AI and democracy. But we should be wary of letting our thinking be guided by such scenarios. To begin with, imagining a future like this presupposes that for a whole range of issues there is a 'most intelligent' solution that for various reasons we have just been unable to put into practice. But intelligence research does not even acknowledge the conceptual uniqueness of intelligence, that is, that there is only one kind of intelligence.[53] Appeals to pure intelligence are illusionary, and allowing algorithms to engage judgments and decisions like this harbors dangers. It might amount to brainwashing people, with intelligent beings downgraded to responders to stimuli.[54] Moreover, designing such a system inevitably involves unprecedented efforts at building state capacities, which are subject to hijacking and other abuse. We should not forget that at the dawn of the digital era we also find *George Orwell's 1984*.

This Grand Democratic AI Utopia, a grand narrative itself, also triggers the warnings from our four techno-skeptical narratives: *Mumford* would readily see in such designs the next version of the megamachine, *Heidegger* detect yet more inauthenticity, *Marcuse* pillory the potential for yet more social control, and *Ellul* recognize how in this way the state is ever more inextricably intertwined with advancements of technique.

V. AI AND DEMOCRACY: POSSIBILITIES AND CHALLENGES FOR THE DIGITAL CENTURY

We saw in Section III that modern democracy requires technology to solve its legitimacy problems. Careful design of the materiality of democracy is needed to solve the distant-state and overbearing-executive problems. At the same time, autocracy benefits from technological advances because they make bureaucracies more effective. The grand techno-skeptical narratives add warnings to the prospect of harnessing technology for democratic purposes, which, however, do not undermine efforts to harness technology to advance democracy. Nor should we be guided by any Grand Democratic AI Utopia. What then are the possibilities and challenges of AI for democracy in this digital century? Specifically, how should AI be designed to harness the public sphere, political power, and economic power for democratic purposes, and thus make them akin to *Winner's* inclusive traffic infrastructure?

1. *Public Spheres*

Public spheres are actor-networks to spread and receive information or opinions about matters of shared concern beyond family and friendship ties.[55] Prior to the invention of writing, public spheres were limited to people talking. Their flourishing depended on availability of places

[53] See e.g., HE Gardner, *Frames of Mind: The Theory of Multiple Intelligences* (2011).

[54] On this, see also D Helbing and others, 'Will Democracy Survive Big Data and Artificial Intelligence?' (*Scientific American*, 25 February 2017) www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/.

[55] For a classic study of the emergence of public spheres, see J Habermas, *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society* (1991). For how information spread in different periods, see

where they could do so safely. The printing press mechanized exchange networks, dramatically lowering costs of disseminating information or ideas. Eventually, newspapers became so central to public spheres that the press and later the media collectively were called 'the fourth estate'.[56] After the newspapers and other printed media, there was the telegraph, then radio, film production, and television. Eventually, leading twentieth century media scholars coined slogans to capture the importance of media for contemporary life, most distinctly *Marshall McLuhan* announcing 'the medium is the message' and *Friedrich Kittler* stating 'the media determine our situation'.[57]

'Fourth estate' is an instructive term. It highlights the relevance of the media, and the deference for the more prominent among them, as well as for particular journalists whose voices carry weight with the public. But the term also highlights that media have class interests of sorts: aside from legal regulations, journalists had demographic and educational backgrounds that generated certain agendas rather than others. The ascent of social media, enabled by the Internet, profoundly altered this situation, creating a public sphere where availability of information and viewpoints was no longer limited by 'the fourth estate'. Big Tech companies have essentially undermined the point of referring to media that way.

In the Western world, Google became dominant in internet searches. Facebook, Twitter, and YouTube offered platforms for direct exchanges among individuals and associations at a scale previously impossible. Political theorist *Archon Fung* refers to the kind of democracy that arose this way as 'wide aperture, low deference democracy': a much wider range of ideas and policies are explored than before, with traditional leaders in politics, media, or culture no longer treated with deference but ignored or distrusted.[58] Not only did social media generate new possibilities for networking but also an abundance of data gathered and analyzed to predict trends or target people with messages for which data mining deems them receptive. The 2018 Cambridge Analytica scandal – a British consulting firm obtaining personal data of millions of Facebook users without consent, to be used for political advertising – revealed the potential of data mining, especially for locations where elections tend to be won by small margins.[59]

Digital media have by now generated an online communications infrastructure that forms an important part of the public sphere, whose size and importance will only increase. This infrastructure consists of the paraphernalia and systems that make our digital lives happen, from the hardware of the Internet to institutions that control domain names and the software that maintains the functionality of the Internet and provides tools to make digital spaces usable (browsers, search engines, app stores, etc.). Today, private interests dominate our digital infrastructure. Typically, engineers and entrepreneurs ponder market needs, profiting from the fact that more and more of our lives unfolds on platforms optimized for clicks and virality.

A Blair and others, *Information: A Historical Companion* (2021). For the development of media in recent centuries, see P Starr, *The Creation of the Media: Political Origins of Modern Communications* (2005).

[56] This term has been attributed to Edmund Burke, and thus goes back to a time decades before media played that kind of role in the American version of modern democracy, see J Schultz, *Reviving the Fourth Estate: Democracy, Accountability and the Media* (1998) 49.

[57] M McLuhan, *Understanding Media: The Extensions of Man* (1994); FA Kittler, *Gramophone, Film, Typewriter* (1999).

[58] For the emergence of digital media and their role for democracy, see Fung and Cohen, 'Democracy and the Digital Public Sphere' (n 23). For the formulation I attributed to Fung, see for instance this podcast PolicyCast, '211 Post-expert Democracy: Why Nobody Trusts Elites Anymore' (*Harvard Kennedy School*, 3 February 2020) www.hks.harvard.edu/more/policycast/post-expert-democracy-why-nobody-trusts-elites-anymore.

[59] A Jungherr, G Rivero and D Gayo-Avello, *Retooling Politics: How Digital Media Are Shaping Democracy* (2020) Chapter 9; C Véliz, *Privacy Is Power: Why and How You Should Take Back Control of Your Data* (2021) Chapter 3 (hereafter Véliz, *Privacy Is Power*).

Especially, news is presented to appeal to certain users, which not only creates echo-chambers but spreads a plethora of deliberate falsehoods (disinformation, rather than misinformation) to reinforce the worldviews of those users. Political scientists have long lamented the ignorance of democratic citizens and the resulting poor quality of public decision making.[60] Even well-informed, engaged voters choose based on social identities and partisan loyalties.[61] Digital media reinforce these tendencies. Twitter, Facebook, YouTube, and competitors seek growth and revenue. Attention-grabbing algorithms of social media platforms can sow confusion, ignorance, prejudice, and chaos. AI tools then readily create artificial unintelligence.[62]

Having a public sphere where viewpoints can be articulated with authority recently became much harder through the emergence of deepfakes. Bringing photoshopping to video, deepfakes replace people in existing videos with someone else's likeness. Currently their reach is mostly limited to pornography, but their potential goes considerably beyond that. In recent decades, video has played a distinguished role in inquiry. What was captured on film served as indisputable evidence, in ways photography no longer could after manipulation techniques became widespread. Until the arrival of deepfakes, videos offered an 'epistemic backstop' in contested testimony.[63] Alongside other synthetic media and fake news, deepfakes might help create no-trust societies where people no longer bother to separate truth from falsehood, and no media help them do so.

What is needed to countermand such tendencies is the creation of what media scholar *Ethan Zuckerman* calls 'digital public infrastructure'.[64] Digital public infrastructure lets us engage in public and civic life in digital spaces, with norms and affordances designed around civic values. Figuratively speaking, designing digital public infrastructure is like creating parks and libraries for the Internet. They are devised to inform us, structured to connect us to both people we agree with and people we disagree with, and encourage dialogue rather than simply reinforcing perceptions. As part of the design of such infrastructures, synthetic media must be integrated appropriately, in ways that require clear signaling of how they are put together. People would operate within such infrastructures also in ways that protect their entitlements as knowers and knowns, their epistemic rights.[65]

One option is to create a fleet of localized, community-specific, public-serving institutions to serve the functions in digital space that community institutions have met for centuries in physical places. There must be some governance model, so this fleet serves the public. Wikipedia's system of many editors and authors or Taiwan's digital democracy platform provide

[60] J Brennan, *Against Democracy* (2017); B Caplan, *The Myth of the Rational Voter: Why Democracies Choose Bad Policies* (2nd ed., 2008); I Somin, *Democracy and Political Ignorance: Why Smaller Government Is Smarter* (2nd ed., 2016).

[61] CH Achen and LM Bartels, *Democracy for Realists: Why Elections Do Not Produce Responsive Government* (2017).

[62] M Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (2019) (hereafter Broussard, *Artificial Unintelligence*).

[63] R Rini, 'Deepfakes and the Epistemic Backstop' (2020) 20 *Philosophers' Imprint* 1. See also C Kerner and M Risse, 'Beyond Porn and Discreditation: Promises and Perils of Deepfake Technology in Digital Lifeworlds' (2021) 8(1) *Moral Philosophy and Politics* 81.

[64] For E Zuckerman's work, see E Zuckerman, 'What Is Digital Public Infrastructure' (*Center for Journalism & Liberty*, 17 November 2020) www.journalismliberty.org/publications/what-is-digital-public-infrastructure#_edn3; and E Zuckerman, 'The Case of Digital Public Infrastructure' (*Knight First Amendment Institute*, 17 January 2020) https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure; see also E Pariser and D Allen, 'To Thrive Our Democracy Needs Digital Public Infrastructure'(*Politico*, 1 May 2021) www.politico.com/news/agenda/2021/01/05/to-thrive-our-democracy-needs-digital-public-infrastructure-455061.

[65] S Zuboff, 'The Coup We Are Not Talking About' (*New York Times*, 29 January 2021) www.nytimes.com/2021/01/29/opinion/sunday/facebook-surveillance-society-technology.html; M Risse, 'The Fourth Generation of Human Rights: Epistemic Rights in Digital Lifeworlds' (2021) *Moral Philosophy and Politics* https://doi.org/10.1515/mopp-2020-0039.

inspiring models for decentralized participatory governance.[66] Alternatively, governments could create publicly funded nonprofit corporations to manage and maintain the public's interest in digital life. Specialized AI would be front and center to such work. Properly designed digital public infrastructures would be like *Winner's* inclusive traffic infrastructures and could greatly help solve the distant-state and overbearing-executive problems.

2. *Political Power*

As far as use of AI for maintenance of power is concerned, the Chinese social credit system – a broadly-based system for gathering information about individuals and bringing that information to bear on what people may do – illustrates how autocratic regimes avail themselves of technological advances.[67] Across the world, cyberspace has become a frequent battleground between excessively profit-seeking or outright criminal activities and overly strong state reactions to them. By now many tools exists that help governments rein in such activities, but those same tools then also help authoritarians oppress political activities.[68] While most mass protests in recent years, from Hong Kong to Algeria and Lebanon, were inspired by hashtags, coordinated through social networks and convened by smartphones, governments have learned how to countermand such movements. They control online spaces by blocking platforms and disrupting the Internet.[69]

In his 1961 farewell speech, US president *Dwight D. Eisenhower* famously warned against acquisition of unwarranted influence 'by the military-industrial complex' and against public policy becoming 'captive of a scientific-technological elite'.[70] Those interconnected dangers would be incompatible with a flourishing democracy. *Eisenhower* spoke only years after the Office of Naval Research had partly funded the first Summer Research Project on AI at Dartmouth in 1956, and thereby indicated that the military-industrial complex had a stake in this technology developed by the scientific-technological elite.[71] Decades later, the 2013 Snowden revelations showed what US intelligence could do with tools we can readily classify as specialized AI. Phones, social media platforms, email, and browsers serve as data sources for the state. Analyzing meta-data (who moved where, connected to whom, read what) provides insights into operations of groups and activities of individuals. Private-sector partnerships have considerably enhanced capacities of law enforcement and military to track people (also involving facial, gait, and voice recognition), from illegal immigrants at risk of deportation to enemies targeted for killing.[72]

Where AI systems are deployed as part of the welfare state, they often surveil people and restrict access to resources, rather than providing greater support.[73] Secret databases and little-known AI applications have had harmful effects in finance, business, education, and politics.

[66] On Taiwan, see A Leonard, 'How Taiwan's Unlikely Digital Minister Hacked the Pandemic' (*Wired*, 23 July 2020) www.wired.com/story/how-taiwans-unlikely-digital-minister-hacked-the-pandemic/.

[67] For a recent take, see J Reilly, M Lyu, and M Robertson 'China's Social Credit System: Speculation vs. Reality' (*The Diplomat*, 30 March 2021) https://thediplomat.com/2021/03/chinas-social-credit-system-speculation-vs-reality/. See also B Dickson, *The Party and the People: Chinese Politics in the 21st Century* (2021).

[68] RJ Deibert, *Black Code: Surveillance, Privacy, and the Dark Side of the Internet* (2013); RJ Deibert, *Reset: Reclaiming the Internet for Civil Society* (2020).

[69] Fung and Cohen, 'Democracy and the Digital Public Sphere' (n 23).

[70] For the speech, see DD Eisenhower, 'Farewell Address' (1961) www.ourdocuments.gov/doc.php?flash=false&doc=90&page=transcript.

[71] Crawford, *Atlas of AI* (n 5) 184; Obviously in 1961, AI is not what Eisenhower had in mind.

[72] Crawford, *Atlas of AI* (n 5) Chapter 6. See also Véliz, *Privacy Is Power* (n 59).

[73] V Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (2018).

AI-based decisions on parole, mortgages, or job applications are often opaque and biased in ways that are hard to detect. Democratic ideals require reasons and explanations, but the widespread use of opaque and biased algorithms has prompted one observer to call societies that make excessive use of algorithms 'black-box societies'.[74] If algorithms do things humans find hard to assess, it is unclear what would even count as relevant explanations. Such practices readily perpetuate past injustice. After all, data inevitably reflect how people have been faring *so far*. Thus, they reflect the biases, including racial biases, that have structured exercises of power.[75] Decades ago *Donna Haraway's* 'Cyborg Manifesto', a classic at the intersection of feminist thought and the philosophy of technology, warned the digital age might sustain white capitalist patriarchy with 'informatics of domination'.[76]

Of course, digital technologies can strengthen democracy. In 2011, Iceland produced the first-ever 'crowdsourced' constitutional proposal in the world. In Taiwan, negotiations among authorities, citizens, and companies like Uber and Airbnb were aided by an innovative digital process for deliberative governance called *vTaiwan*. France relied on digital technologies for the Great National Debate in early 2019 and the following Convention on Climate Change between October 2019 and June 2020, experimenting with deliberation at the scale of a large nation.[77] Barcelona has become a global leader in the smart city movement, deploying digital technology for matters of municipal governance,[78] though it is Singapore, Helsinki, and Zurich that do best on the *Smart City Index 2020* (speaking to the fact of how much innovation goes on in that domain).[79] An Australian, non-profit, eDemocracy project, openforum.com.au, invites politicians, senior administrators, academics, businesspeople, and other stakeholders to engage in policy debates. The California Report Card is a mobile-optimized web application promoting public involvement in state government. As the COVID-19 pandemic ravaged the world, democracies availed themselves of digital technologies to let people remain connected and serve as key components of public health surveillance. And while civil society organizations frequently are no match for abusive state power, there are remarkable examples of how even investigations limited to open internet sources can harvest the abundance of available data to pillory abuse of power. The best-known example is the British investigative journalism website Bellingcat that specializes in fact-checking and open-source intelligence.[80]

One striking fact specifically about the American version of modern democracy is that, when preferences of low- or middle-income Americans diverge from those of the affluent, there is virtually no correlation between policy outcomes and desires of the less advantaged groups.[81]

[74] F Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (2016). See also Broussard, *Artificial Unintelligence* (n 62); C O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2017).

[75] R Benjamin, *Race After Technology Race After Technology: Abolitionist Tools for the New Jim Code* (2019); R Benjamin, *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life* (2019); SU Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018). See also C D'Ignazio and LF Klein, *Data Feminism* (2020); S Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need* (2020).

[76] D Haraway, *Simians, Cyborgs, and Women: The Reinvention of Nature* (2015) 149–182.

[77] L Bernholz, H Landemore, and R Reich, *Digital Technology and Democratic Theory* (2021).

[78] P Preville, 'How Barcelona Is Leading a New Era of Digital Democracy' (*Medium*, 13 November 2019) https://medium.com/sidewalk-talk/how-barcelona-is-leading-a-new-era-of-digital-democracy-4a033a98cf32.

[79] IMD, 'Smart City Index' (*IMD*, 2020) www.imd.org/smart-city-observatory/smart-city-index/.

[80] E Higgins, *We Are Bellingcat: Global Crime, Online Sleuths, and the Bold Future of News* (2021). See also M Webb, *Coding Democracy: How Hackers Are Disrupting Power, Surveillance, and Authoritarianism* (2020).

[81] LM Bartels, *Unequal Democracy: The Political Economy of the New Gilded Age* (2018); M Gilens, *Affluence and Influence: Economic Inequality and Political Power in America* (2014).

As far as political power is concerned, the legitimacy of modern democracy is questionable indeed. Democracy could be strengthened considerably by well-designed AI. Analyzing databases would give politicians a more precise image of what citizens need. The bandwidth of communication between voters and politicians could increase immensely. Some forms of surveillance will be necessary, but democratic governance requires appropriate oversight. The digital public infrastructure discussed in the context of the public sphere can be enriched to include systems that deploy AI for improving citizen services. The relevant know-how exists.[82]

3. *Economic Power*

Contemporary ideals of democracy include egalitarian empowerment of sorts. But economic inequality threatens any such empowerment. Contemporary democracies typically have capitalist economies. As French economist *Thomas Piketty* has argued, over time capitalism generates inequality because, roughly speaking, owners of shares of the economy benefit more from it than people living on the wages the owners willingly pay.[83] A worry about democracy across history (also much on the mind of Publius) has been that masses would expropriate elites. But in capitalist democracies, we must worry about the opposite. It takes sustained policies around taxation, transportation, design of cities, healthcare, digital infrastructure, pension and education systems, and macro-economic and monetary policies to curtail inequality.

One concern about AI is that, generally, the ability to produce or use technology is one mechanism that drives inequality, enabling those with requisite skills to advance – which enables them not only to become well-off but to become owners in the economy in ways that resonate across generations. Technology generally and AI specifically are integral parts of the inequality-enhancing mechanisms *Piketty* identifies. One question is how the inequality-increasing tendencies play out for those who are not among the clear winners. AI will profoundly transform jobs, at least because aspects of many jobs will be absorbed by AI or otherwise mechanized. These changes also create new jobs, including at the lower end, in the maintenance of the hardware and the basic tasks around data gathering and analysis.[84] On the optimistic side of predictions about the future of work, we find visions of society with traditional jobs gradually transformed, some eliminated and new jobs added – in ways that create much more leisure time for the average people, owing to increased societal wealth.

On the pessimistic side, many who are unqualified for meaningful roles in tech economies might be dispensable to the labor force. Their political relevance might eventually amount to little more than that they must be pacified if they cannot be excluded outright. Lest this standpoint be dismissed as Luddite alarmism ('at the end of the tunnel, there have always been more jobs than before'), we should note that economies where data ownership becomes increasingly relevant and AI absorbs many tasks could differ critically from economies organized around ownership of land or factories. In those earlier cases large numbers of people were needed to provide labor, in the latter case also as consumers. Elites could not risk losing too many laborers. But this constraint might not apply in the future. To be sure, a lot here will

[82] On AI and citizen services, see H Mehr, 'Artificial Intelligence for Citizen Services and Government' (*Harvard Ash Center Technology & Democracy Fellow*, August 2017) https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf.

[83] T Piketty, *Capital in the Twenty First Century* (2014).

[84] On these topics, see e.g., D Susskind, *A World Without Work: Technology, Automation, and How We Should Respond* (2020); DM West, *The Future of Work: Robots, AI, and Automation* (2019).

depend on how questions around control over, or ownership of, data are resolved, questions whose relevance for our future economy cannot be overstated.[85]

As recently argued by *Shoshana Zuboff*, the importance of data collection for the economy has become so immense that the term 'surveillance capitalism' characterizes the current stage of capitalism.[86] Surveillance capitalism as an economic model was developed by Google, which to surveillance capitalism is what Ford was to mass production. Later, the model was adopted by Facebook, Amazon, and others. Previously, data were collected largely to improve services. But subsequently, data generated as byproducts of interactions with multifarious devices were deployed to develop predictive products, designed to forecast what we will feel, think, or do, but ultimately also to control and change it, always for the sake of monetization. *Karl Marx* and *Friedrich Engels* identified increasing commodification as a basic mechanism of capitalism (though they did not use that very term). Large-scale data collection is its maximal version: It commodifies all our lived reality.

In the twentieth century, *Hannah Arendt* and others diagnosed mechanisms of 'totalitarian' power, the state's all-encompassing power.[87] Its central metaphor is Big Brother, capturing the state's omnipresence. Parallel to that, *Zuboff* talks about 'instrumentarian' power, exercised through use of electronic devices in social settings for harvesting profits. The central metaphor is the 'Big Other', the ever-present electronic device that knows just what to do. Big Brother aimed for total control, Big Other for predictive certainty.

Current changes are driven by relatively few companies, which futurist *Amy Webb* calls 'the Big Nine': in the US, Google, Microsoft, Amazon, Facebook, IBM and Apple, in China Tencent, Alibaba and Baidu.[88] At least at the time of *Webb's* writing, the Chinese companies were busy consolidating and mining massive amounts of data to serve the government's ambitions; the American ones implemented surveillance capitalism, embedded into a legal and political framework that, as of 2021, shows little interest in developing strategic plans for a democratic future and thus do for democracy what the Chinese Communist party did for its system – upgrade it into this century. To be sure, the EU is more involved in such efforts. But none of the Big Nine are based there, and overall, the economic competition in the tech sector seems to be ever more between the United States and China.

The optimistic side of predictions about the future of work seems reachable. But to make that happen in ways that also strengthen democracy, both civil society and the state must step up, and the enormous power concentrated in Big Tech companies needs to be harnessed for democratic purposes.

VI. CONCLUSION

Eventually there might be a full-fledged Life 3.0, whose participants not only design their cultural context (as in Life 2.0, which sprang from the evolutionary Life 1.0), but also their physical shapes.[89] Life 3.0 might be populated by genetically enhanced humans, cyborgs,

[85] On this, also see M Risse, 'Data as Collectively Generated Patterns: Making Sense of Data Ownership' (*Carr Center for Human Rights Policy*, 4 April 2021) https://carrcenter.hks.harvard.edu/publications/data-collectively-generated-patterns-making-sense-data-ownership.

[86] S Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (2019). See also Véliz, *Privacy Is Power* (n 59).

[87] H Arendt, *The Origins of Totalitarianism* (1973).

[88] A Webb, *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity* (2020).

[89] For that term, see Tegmark, *Life 3.0* (n 3).

uploaded brains, as well as advanced algorithms embedded into any manner of physical device. If Life 3.0 ever emerges, new questions for governance arise. Would humans still exercise control? If so, would there be democracies, or would some people or countries subjugate everybody else? Would it be appropriate to involve new intelligent entities in governance, and what do they have to be like for the answer to be affirmative? If humans are not in control, what will governance be like? Would humans even be involved?[90]

It is unclear when questions about democracy in Life 3.0 will become urgent. Meanwhile, as innovation keeps happening, societies will change. Innovation will increase awareness of human limitations and set in motion different ways for people to deal with them. As *Norbert Wiener*, whose invention of cybernetics inaugurated later work on AI, stated in 1964:

> The world of the future will be an ever more demanding struggle against the limitation of our intelligence, not a comfortable hammock in which we can lie down to be waited upon by our robot slaves.[91]

Maybe more and more individuals will want to adapt to technological change and perhaps deploy technology to morph into a transhuman stage.[92] Generally, what technology they use – the materiality of their lives – affects who people are and want to become. Technology mediates how we see ourselves in relation to our social and natural environment, how we engage with people, animals, and material objects, what we do with ourselves, how we spend our professional lives, etc. In short, technology critically bears on what forms of human life get realized or even imagined. For those that do get realized or imagined, what it means to be part of them cannot be grasped without comprehending the role of technology in them.[93]

As we bring about the future, computer scientists will become ever more important, also as experts in designing specialized AI for democratic purposes. That raises its own challenges. Much as technology and democracy are no natural allies, technologists are no natural champions of, nor even qualified advisers to, democracy. Any scientific activity, as *Arendt* stated some years before *Wiener* wrote the words just cited, as it acts into nature from the standpoint of the universe and not into the web of human relationships, lacks the revelatory character of action as well as the ability to produce stories and become historical, which together form the very source from which meaningfulness springs into and illuminates human existence.[94]

Democracy is a way of life more than anything else, one that greatly benefits from the kind of action *Arendt* mentions. And yet modern democracy critically depends on technology to be the kind of actor-network that solves the distant-state and overbearing-executive problems. Without suitable technology, modern democracy cannot survive. Technology needs to be consciously harnessed to become like *Winner's* inclusive traffic infrastructure, and both technologists and citizens generally need to engage with ethics and political thought to have the spirit and dedication to build and maintain that kind of infrastructure.

[90] Tegmark, Chapter 5 (n 3).

[91] N Wiener, *God and Golem, Inc.; a Comment on Certain Points Where Cybernetics Impinges on Religion* (1964) 69.

[92] D Livingstone, *Transhumanism: The History of a Dangerous Idea* (2015); M More and N Vita-More, *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future* (2013); Bostrom, *Superintelligence* (n 3).

[93] For some aspects of this, NC Carr, *The Shallows: How the Internet Is Changing the Way We Think, Read and Remember* (2011); S Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other* (2017). But the constitutive role of technology on human life is a central theme in the philosophy of technology and adjacent areas generally.

[94] H Arendt, *The Human Condition* (1958) 324.

7

The New Regulation of the European Union on Artificial Intelligence

Fuzzy Ethics Diffuse into Domestic Law and Sideline International Law

Thomas Burri

I. INTRODUCTION

In the conventional picture, international law emanates from treaties states conclude or customs they observe. States comply with binding international law and ensure compliance in the domestic context. In this picture, states in a 'top-top' process agree on the law before it trickles down to the domestic legal order where it is implemented. Norms made in other ways are considered 'soft', which implies that they provide mere guidance but are technically not binding, or irrelevant to international law.

Obviously, there is room for nuance in the conventional take on international law and its sources. Soft law, for instance, can acquire authority that comes close to binding character.[1] It can also serve to interpret binding law that would otherwise remain ambiguous.[2] However, traditional international law ignores that law is also created outside of its formal processes. Norms can notably consolidate independently from the will of states in speedy, subcutaneous processes. Norms can diffuse subliminally across the world into municipal laws which incorporate and make them binding domestically. In this informal process, international law enters the stage late, if at all. It can only retrace the law that has already been locked in domestically. This informal process resembles 'bottom-up international law',[3] though its character is more 'bottom to bottom' and 'transnational'. The process shall be referred to as 'norm diffusion' in this chapter. It is illustrated through the creation of norms governing Artificial Intelligence (AI).

The informal process of law creation described above is far from ubiquitous. It can be hard to trace, for when international law codifies or crystallizes 'new' norms, it tends to obscure their origin in previous processes of law creation. It is also messy, for it does not adhere to the hierarchies that distinguish conventional international law. Even more so, it is worth discussing norm diffusion to complement the picture of international law and its sources.

The present chapter could have examined norm diffusion in the current global public health crisis. It seems that in the COVID-19 pandemic, behavioural norms informed by scientific

[1] See the treatment accorded to the ICJ, *Legal Consequences of the Separation of the Chagos Archipelago from Mauritius in 1965, Advisory Opinion* [2019] ICJ Rep 95, in *Dispute concerning Delimitation of the Maritime Boundary between Mauritius and Maldives in the Indian Ocean, no. 28 (Mauritius/Maldives) (Preliminary Objections)* ITLOS (2021) para. 203; see our discussion in T Burri and J Trinidad, 'Introductory note' (2021) 60(6) International Legal Materials 969–1037.

[2] Article 32 Vienna Convention on the Law of the Treaties, 1155 UNTS 331 (engl.) 27 March 1969.

[3] J Koven Levit, 'Bottom-Up International Lawmaking: Reflections on the New Haven School of International Law' (2007) 32 *The Yale Journal of International Law* 393–420.

expertise take shape rapidly, diffuse globally, and are incorporated into domestic law. In contrast, international lawyers are only now beginning to discuss a more suitable legal framework. However, rather than engaging with the ongoing chaotic normative process in public health, this chapter discusses a more mature and traceable occurrence of norm diffusion, namely that of the regulation of AI. The European Commission's long-awaited proposal from April 2021 for a regulation on AI marks the perfect occasion to illustrate the diffusion of AI norms.

This chapter proceeds in three steps. First, it examines the creation of ethical norms designed to govern AI (Section II). Second, it investigates the diffusion of such norms into domestic law (Section III). This section examines the European Commission's recent legislative proposal to show how it absorbs ethical norms on AI. This examination likewise sheds light on the substance of AI norms. Section III could also be read on its own, in other words, without regard to international law-making, if one wished to learn only about the origins and the substance of the European Union regulation in the offing. Section IV then discusses how the process of norm diffusion described in Sections II and III sidelines international law. Section V concludes and offers an outlook.

II. THE CREATION OF ETHICAL NORMS ON AI

The creation of ethical norms governing AI has taken many forms over a short period of time. It began with robotics. Roughly 50 years ago, *Isaac Asimov's* science fiction showed how ambiguous certain ethical axioms were when applied to intelligent robots.[4] Since then, robotics has made so much progress that scientists have begun to take an interest in ethical principles for robotics. Such principles, which were prominently enunciated in the United Kingdom in 2010, addressed the potential harm caused by robots, responsibility for damage, fundamental rights in the context of robotics, and several other topics, including safety/security, deception, and transparency.[5] The same or similar aspects turned out to be relevant for AI after it had re-awakened from hibernation. Two initiatives were significant in this regard, namely the launch of the One Hundred Year Study on Artificial Intelligence at Stanford University in 2014[6] and an

[4] See A Winfield, 'An Updated Round Up of Ethical Principles of Robotics and AI' (*Alan Winfield's Web Log*, 18 April 2019) https://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html: '1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.' The work of the present author has benefitted tremendously from Winfield's collation of ethics principles on AI in his blog at a time when it was not yet easy to assemble the various sets of ethics principles. For the primary source of Asimov's principles, see e.g. I Asimov, *The Caves of Steel* (1954) and I Asimov, *The Naked Sun* (1957); for a discussion of Asimov's principles about fifty years after Asimov had begun writing about them, see RR Murphy and DD Woods, 'Beyond Asimov: The Three Laws of Responsible Robotics' (2009) July/August 2019 *IEEE Intelligent Systems* 14–20.

[5] Drafted in the context of the Engineering and Physical Sciences Research Council and the Arts and Humanities Research Council (United Kingdom) in 2010, but published only in M Boden and others, 'Principles of Robotics: Regulating Robots in the Real World' (2017) 29 *Connection Science* (2) 124–129; see also A Winfield, 'Roboethics – for Humans' (2011) 17 May 2011 *The New Scientist* 32–33. Before that, ethicists and philosophers had already discussed robotics in various perspectives, see e.g. R Sparrow, 'Killer Robots' (2007) 24 *Journal of Applied Philosophy* (1) 62–77, RC Arkin, *Governing Lethal Behavior in Autonomous Robots* (2009); PW Singer, *Wired for War: The Robotics Revolution and Conflict in the Twenty-First Century* (2009); W Wallach and C Allen, *Moral Machines: Teaching Robots Right from Wrong* (2009).

[6] See E Horvitz, *One Hundred Year Study on Artificial Intelligence: Reflections and Framing* (2014) https://ai100.stanford.edu/reflections-and-framing (hereafter Horvitz, 'One Hundred Year Study') also for the roots of this study (on p 1).

Open Letter[7] signed by researchers and entrepreneurs in 2015.[8] Both initiatives sought to guide research toward beneficial and robust AI.[9] In their wake, the IEEE, an organization of professional engineers, in 2015 embarked on a broad public initiative aimed at pinning down the ethics of autonomous systems;[10] a group of AI professionals gathered to generate the Asilomar principles for AI, which were published in 2017[11]; and an association of experts put forward ethical principles for algorithms and programming.[12] This push to establish ethical norms occurred in

[7] Future of Life Institute, 'An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence' (*Future of Life Institute*) http://futureoflife.org/ai-open-letter/ (hereafter 'Open Letter'); another important moment before the Open Letter was a newspaper article: S Hawking and others, 'Transcendence Looks at the Implications of Artificial Intelligence – But Are We Taking AI Seriously Enough?' *The Independent* (1 May 2014) www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html.

[8] Several research groups had addressed the law and ethics of robots in the meanwhile: see C Leroux and others, 'Suggestion for a Green Paper on Legal Issues in Robotics' (31 December 2012) www.researchgate.net/publication/310167745_A_green_paper_on_legal_issues_in_robotics; E Palmerini and others, 'Guidelines on Regulating Robotics' (*Robo Law*, 22 September 2014) www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf; other authors previously had prepared the ground, notably P Lin, K Abney, and GA Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (2012); U Pagallo, *The Law of Robots: Crimes, Contracts, Torts* (2013); N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (2014) (hereafter Bostrom, 'Superintelligence'); JF Weaver, *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws* (2014); M Anderson and S Anderson Leigh, 'Towards Ensuring Ethical Behaviour from Autonomous Systems: A Case-Supported Principle-Based Paradigm' (2015) 42 *Industrial Robot: An International Journal* (4) 324–331.

[9] In the 100 Year Study, law and ethics figured prominently as a research topic (Horvitz, 'One Hundred Year Study'(n 6) topics 6 and 7), while the Open Letter (n 7) included a research agenda parts of which were 'law' and 'ethics'.

[10] The first version of 'Ethically Aligned Design' was made public in 2016: Institute of Electrical and Electronics Engineers (IEEE), 'Ethically Aligned Design, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems' (13 December 2016) http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf; meanwhile, a first edition has become available: Institute of Electrical and Electronics Engineer (IEEE), 'Ethically Aligned Design, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems' (2019) https://ethicsinaction.ieee.org; in the following, reference is made to the latter, the first edition (hereafter, IEEE, 'Ethically Aligned Design'). It contains a section on high-level 'general principles' which address human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence. Other sections of the Charter discuss classical ethics, well-being, affective computing, personal data and individual agency, methods to guide ethical research and design, sustainable development, embedding values, policy, and law. The last section on the 'law' focuses on fostering trust in autonomous and intelligent systems and the legal status of such systems. For full disclosure, the present author co-authored the section on law of Ethically Aligned Design.

[11] Future of Life Institute, 'Asilomar AI Principles' (*Future of Life Institute*, 2017) https://futureoflife.org/ai-principles/ (hereafter Future of Life Institute, 'Asilomar AI Principles'). The Asilomar principles address AI under three themes, namely 'research', 'ethics and values', and 'longer term issues'. Several sub-topics are grouped under each theme, viz. goal, funding, science-policy link, culture, race avoidance (under 'research'); safety, failure transparency, judicial transparency, responsibility, value alignment, human values, personal privacy, liberty and privacy, shared benefit, shared prosperity, human control, non-subversion, arms race (under 'ethics and values'); and capability caution, importance, risks, recursive self-improvement, and common good (under 'longer term issues').

[12] Association for Computing Machinery US Public Policy Council (USACM), 'Statement on Algorithmic Transparency and Accountability' (*USACM*, 12 January 2017) www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf (hereafter USACM, 'Algorithmic Transparency'); the principles are part of a broader code of ethics: Association for Computing Machinery Committee on Professional Ethics, 'ACM Code of Ethics and Professional Conduct' (*ACM Ethics*, 22 June 2018) https://ethics.acm.org. Summed up, the principles are the following: 1. Be aware of bias; 2. Enable questioning and redress; 3. If you use algorithms, you are responsible even if not able to explain; 4. Produce explanations; 5. Describe the data collection process, while access may be restricted; 6. Record to enable audits; 7. Rigorously validate your model and make the test public. Compare also with the principles a professional organization outside of the anglophone sphere published relatively early: Japanese Society for Artificial Intelligence, 'The Japanese Society for Artificial Intelligence Ethical Guidelines' (2017) http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf (hereafter Japanese Society for AI, 'Guidelines') in summary: 1. Contribute to humanity, respect human rights and diversity, eliminate threats to safety; 2. Abide by the law, do not use AI to harm others, directly or indirectly; 3. Respect privacy; 4. AI as a resource is to be used fairly and equally by humanity, avoid discrimination and inequality; 5. Be sure to maintain AI safe and under control; provide users with

lockstep with the significant technological advances in AI,[13] and it is against this background that it must be understood.

In parallel, a discussion began to take shape within the Convention on Certain Conventional Weapons (CCW)[14] in Geneva. This discussion soon shifted its focus to the use of force by means of autonomous systems.[15] It notably zeroed in on physically embodied weapons systems – a highly specialized type of robot – and refrained from considering disembodied weapons, sometimes called cyberweapons.[16] The focus on embodiment[17] had the effect of keeping AI out of the limelight in Geneva for a long time.[18] As a broader consequence, the international law

appropriate and sufficient information; 6. Act with integrity and so that society can trust you; 7. Verify performance and impact of AI, warn if necessary, prevent misuse; whistle blowers shall not be punished; 8. Improve society's understanding of AI, maintain consistent and effective communication; 9. Have AI abide by these guidelines in order for it to become a quasi-member of society. Note, in particular, the Japanese twist of the last guideline.

[13] See by way of example V Mnih, and others, 'Human-Level Control through Deep Reinforcement Learning' (2015) 518 *Nature* (26 February 2015) 529–533; see also B Schölkopf, 'Learning to See and Act' (2015) 518 *Nature* (26 February 2015) 486–487; and D Silver and others 'Mastering the Game of Go with Deep Neural Networks and Tree Search' (2016) 529 *Nature* (28 January 2016) 484–489. The Darpa Challenges also significantly pushed research forward, see T Burri, 'The Politics of Robot Autonomy' (2016) 7 *European Journal of Risk Regulation* (2) 341–360. In robotics, a certain amount of hysteria has been created by Boston Dynamics' videos. An early example is the video about the Atlas robot: Boston Dynamics, 'Atlas, the Next Generation' (YouTube, 23 February 2016) www.youtube.com/watch?v=rVlhMGQgDkY&app=desktop. But it is not all hype and hysteria, see already GA Pratt, 'Is a Cambrian Explosion Coming for Robotics?' (2015) 29 *Journal of Economic Perspectives* (3 (Summer 2015)) 51–60.

[14] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (with Protocols I, II, and III), 1342 UNTS 163 (English), 10 October 1980.

[15] This discussion was spurred on by a report: Human Rights Watch and Harvard International Human Rights Clinic, 'Losing Humanity: The Case against Killer Robots' (*HRW*, 19 November 2012) www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots, and an international civil society campaign, the Campaign to Stop Killer Robots (see www.stopkillerrobots.org), in which from the beginning researchers such as P Asaro, R Sparrow, N Sharkey, and others were involved; the International Committee for Robot Arms Control (ICRAC, see www.icrac.net) also campaigned against Killer Robots. Much of the influential legal work within the context of the Campaign goes back to B Docherty, e.g. the report just mentioned or B Docherty, 'Mind the Gap: The Lack of Accountability for Killer Robots' (*HRW*, 9 April 2015) www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots; B Docherty, 'Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition' (*HRW*, 8 November 2015) www.hrw.org/news/2015/11/08/precedent-preemption-ban-blinding-lasers-model-killer-robots-prohibition. The issue of autonomous weapons systems had previously been addressed by Philip Alston: UNCHR, 'Interim Report by UN Special Rapporteur on extrajudicial, summary or arbitrary executions, Philip Alston' (2010) UN Doc A/65/321; see also P Alston, 'Lethal Robotic Technologies: The Implications for Human Rights and International Humanitarian Law' (2011) 21 *Journal of Law, Information and Science* 35-60; and later by Christof Heyns: UNCHR, 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, (2013) UN Doc A/HRC/23/47; for scholarship, see A Leveringhaus, *Ethics and Autonomous Weapons* (2016).

[16] The discussion of cyber warfare took a different path. See most recently, D Trusilo and T Burri, 'Ethical Artificial Intelligence: An Approach to Evaluating Disembodied Autonomous Systems' in R Liivoja and A Väljataga (eds), *Autonomous Cyber Capabilities under International Law* (2021) 51–66 (hereafter Trusilo and Burri, 'Ethical AI').

[17] For a discussion of embodiment from a philosophical perspective, see C Durt, 'The Computation of Bodily, Embodied, and Virtual Reality' (2020) 1 *Phänomenologische Forschungen* 25–39 www.durt.de/publications/bodily-embodied-and-virtual-reality/.

[18] Defence has meanwhile gone beyond autonomy to consider also AI. Contrast the early US Department of Defence, 'Directive on Autonomy in Weapon Systems' (DoD, 21 November 2012, amended 8 May 2017) www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf with the recent Defense Innovation Board, 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense' (DoD, 24 February 2020) 12 https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF : 'The important thing to consider going forward is that however DoD integrates AI into autonomous systems, whether or not they are weapons systems, sharp ethical and technical distinctions between AI and autonomy may begin to blur, and the Department should consider the interaction between AI and autonomy to ensure that legal and ethical dimensions are considered and addressed.' The Report addresses AI within the Department of Defense in general, not just in combat. It posits five key aspects which should inform the Department of Defense's engagement with AI: Responsible, equitable, traceable, reliable, governable. ('Equitable' refers to what is in other documents often

community became fixated on an exclusive and exotic aspect – namely physical ('kinetic') autonomous weapons systems – while the technological development was more comprehensive. Despite their narrow focus, the seven years of discussions in Geneva have yielded few concrete results, other than a great deal of publicity.[19]

At about the same time, autonomous cars also became the subject of ethical discussion. This discussion, however, soon got bogged down in largely theoretical, though fascinating, ethical dilemmas, such as the trolley problem.[20] However, unlike those gathered in Geneva to ponder autonomous weapons systems, those intent on putting autonomous cars on the road were pragmatic. They found ways of generating meaningful output that could be implemented.[21]

In 2017, the broader public beyond academic and professional circles became aware of the promises and perils of AI. Civil society began to discuss the ethics of AI and soon produced tangible output.[22] Actionable principles were also proposed on behalf of

called 'fairness' or 'avoidance of bias', terms which, according to the report, may be misleading in defence, see p 31). See also HM Roff, 'Artificial Intelligence: Power to the People' (2019) 33 *Ethics and International Affairs* 127, 128–133, for a distinction between automation, autonomy, and AI.

[19] The output consists of eleven high-level principles on autonomous weapons systems: Alliance for Multilateralism on Lethal Autonomous Weapons Systems (LAWS), 'Eleven Guiding Principles on Lethal Autonomous Weapons Systems' (*Alliance for Multilateralism*, 2020) https://multilateralism.org/wp-content/uploads/2020/04/declaration-on-lethal-autonomous-weapons-systems-laws.pdf (hereafter Eleven Guiding Principles on Lethal Autonomous Weapons); for the positions of states within CCW and the status quo of the discussions, see D Lewis, 'An Enduring Impasse on Autonomous Weapons' (*Just Security*, 28 September 2020) www.justsecurity.org/72610/an-enduring-impasse-on-autonomous-weapons/; for a thorough discussion of autonomous weapons systems and AI see AL Schuller, 'At the Crossroads of Control: The Intersection of Artificial Intelligence and Autonomous Weapons Systems with International Humanitarian Law' (2017) 8 *Harvard National Security Journal* (2) 379–425; see also SS Hua, 'Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control' (2019) 51 *Georgetown Journal of International Law* 117–146.

[20] See JF Bonnefon, A Shariff, and I Rahwan, 'The Social Dilemma of Autonomous Vehicles' (2016) 352 *Science* (6293) 1573–1576; E Awad and others, 'The Moral Machine Experiment' (2018) 563 *Nature* 59–64.

[21] Note in particular, Ethics Commission of the Federal Ministry of Transport and Digital Infrastructure, 'Automated and Connected Driving' (*BMVI*, June 2017) www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile. This report pinpointed 20 detailed principles. The principles stated clearly that autonomous driving was ethically justified under certain conditions, even if the result of autonomous driving was that persons may occasionally be killed (see principles 2, 8, and 9). See also A von Ungern-Sternberg, 'Autonomous Driving: Regulatory Challenges Raised by Artificial Decision-Making and Tragic Choices' in W Barfield and U Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (2017) 251–278.

[22] The Future Society in Policy Research, The Law & Society Initiative, 'Principles for the Governance of AI' (*The Future Society*, 15 July 2017) https://thefuturesociety.org/the-law-society-initiative/> (under 'learn more'); University of Montreal, 'Montreal Declaration for a Responsible Development of Artificial Intelligence' (*Montréal Declaration Responsible AI_*, 2018) https://docs.wixstatic.com/ugd/ebc3a3_c5c1c196fc164756afb92466c081d7ae.pdf (hereafter 'Montreal Declaration for AI') was one of the first documents to examine the societal implications of AI, putting forward a very broad and largely aspirational set of principles, the gist being: 1. Increase well-being (with 5 sub-principles); 2. Respect people's autonomy and increase their control over lives (6 sub-principles); 3. Protect privacy and intimacy (8); 4. Maintain bonds of solidarity between people and generations (6); 5. Democratic participation in AI: it must be intelligible, justifiable, and accessible, while subject to democratic scrutiny, debate, and control (10); 6. Contribute to just and equitable society (7); 7. Maintain diversity, do not restrict choice and experience (6); 8. Prudence: exercise caution in development, anticipate adverse consequences (5); 9. Do not lessen human responsibility (5); 10. Ensure sustainability of planet (4). Compare with: Amnesty International and Access Now, 'The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems' (16 May 2018) www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf (hereafter 'Toronto Declaration') which, although put together by non-governmental organizations, is more in the nature of an academic legal text and not easily summarized. It emphasizes the duties of states to identify risks, ensure transparency and accountability, enforce oversight, promote equality, and hold the private sector to account. Similar duties are incumbent on private actors, though they are less firm. The right to effective remedy is also emphasized. Compare

women[23] and labour,[24] and *AlgorithmWatch*, a now notable non-governmental organization, was founded.[25]

In step with civil society, private companies adopted ethical principles concerning AI.[26] Such principles took different shapes depending on companies' fields of business. The principles

also with The Public Voice, 'Universal Guidelines for Artificial Intelligence' (*The Public Voice*, 23 October 2018) https://thepublicvoice.org/ai-universal-guidelines/.

[23] Women Leading in AI, '10 Principles of Responsible AI' (*Women Leading in AI*, 2019) https://womenleadinginai.org/wp-content/uploads/2019/02/WLiAI-Report-2019.pdf. This initiative did not look at AI strictly from a gender, but a broader societal perspective. The 10 principles can be summarized as follows: 1. Mirror the regulatory approach for the pharmaceutical sector; 2. Establish an AI regulatory body with powers inter alia to: audit algorithms, investigate complaints, issue fines for breaches of the General Data Protection Regulation, the law and equality, and ensure algorithms are explainable. 3. Introduce 'Certificate of Fairness for AI systems'; 4. Require 'Algorithm Impact Assessment' when AI is employed with impact on individuals; 5. In public sector, inform when decisions are made by machines; 6. Reduce liability when 'Certificate of Fairness' is given; 7. Compel companies to bring their workforce with them; 8. Establish digital skills funds to be fed by companies; 9. Carry out skills audit to identify relevant skills for transition; 10. Establish education and training programme, especially to encourage women and underrepresented sections of society.

[24] UNI Global Union, '10 Principles for Ethical AI, UNI Global Union Future World of Work' (*The Future World of Work*, 2017) www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf; summarized: 1. Transparency; 2. Equip with black box; 3. Serve people/planet; 4. Humans must be in command, incl. responsibility, safety, compliance with privacy and law; 5. Avoid bias in AI; 6. Share benefits; 7. Just transition for workforce and support for human rights; 8. Establish global multi-stakeholder governance mechanism for work and AI; 9. Ban responsibility of robots; 10. Ban autonomous weapons.

[25] See https://algorithmwatch.org/en/transparency/; AlgorithmWatch provides a useful database bringing together ethical guidelines on AI: https://inventory.algorithmwatch.org/. In 2017, the AI Now Institute at New York University, which conducts research on societal aspects of AI, was also established (see www.ainowinstitute.org). Various 'research agendas' have by now been published: J Whittlestone and others, 'Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research' (*Nuffield Foundation*, 2019) www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf (with a useful literature review in appendix 1 and a review of select ethics principles in appendix 2); A Dafoe, 'AI Governance: A Research Agenda' (*Future of Humanity Institute*, 2018) www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf, which broadly focuses on economics and political science research. Compare with OpenAI which is on a 'mission' to ensure that general AI will be beneficial. For this purpose, it conducts research on AI based on its own ethical Charter: OpenAI, OpenAI Charter (*Open AI*, 9 April 2018) https://openai.com/charter/ (hereafter OpenAI Charter); in brief, the principles of the Charter are: Ensure general AI benefits all, avoid uses that harm or concentrate power; primary duty to humanity, minimize conflicts of interest that compromise broad benefit; do the research that makes general AI safe; if late-stage development of general AI becomes a competitive race without time for precaution, stop competing and assist the other project; leadership in technology, policy, and safety advocacy is not enough; AI will impact before general AI, so lead there too; cooperate actively, create global community; provide public goods that help society navigate towards general AI; for now, publish most AI research, but later probably not for safety reasons.

[26] See Intel, 'AI Public Policy Opportunity' (*Intel*, 2017) https://blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf summed up: 1. Foster innovation and open development; 2. Create new human employment and protect people's welfare; 3. Liberate data responsibly; 4. Rethink privacy; 5. Require accountability for ethical design and implementation. Further examples include Sage, 'The Ethics of Code: Developing AI for Business with Five Core Principles' (*Sage*, 2017) www.sage.com/~/media/group/files/business-builders/business-builders-ethics-of-code.pdf?la=en&hash=CB4DF0EB6CCB15F55E72EBB3CD5D526B (hereafter Sage, 'The Ethics of Code'), in brief: 1. Reflect diversity, avoid bias; 2. Accountable AI, but also accountable users; AI must not be too clever to be held accountable; 3. Reward AI for aligning with human values through reinforcement learning; 4. AI should level playing field: democratize access, especially for disabled persons; 5. AI replaces, but must also create work: humans should focus on what they are good at; Google, 'Artificial Intelligence at Google: Our Principles' (*Google*, 2018) https://ai.google/principles/ (hereafter Google, 'AI Principles'); in brief: 1. Be socially beneficial and thoughtfully evaluate when to make technology available on non-commercial basis; 2. Avoid bias; 3. Build and test for safety; 4. Be accountable to people, i.e. offer feedback, explanation, and appeal; subject AI to human direction and control; 5. Incorporate privacy design principles; 6. Uphold high standard of scientific excellence; 7. Use of AI must accord with these principles; 8. No-go areas: technology likely to cause overall harm; weapons; technology for surveillance violating internationally accepted norms; technology whose purpose violates international law and human rights – though this 'point 8' may evolve; IBM, 'Everyday Ethics for Artificial Intelligence' (*IBM*, 2018) www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (hereafter IBM, 'Ethics for AI'); in brief: 1. Be accountable, i.e.

embody a certain degree of self-commitment, which is not subject to outside verification, though.[27] Parts of the private sector and the third sector have also joined forces, most prominently in the Partnership on AI and its tenets on AI.[28]

The development has not come to a halt today. Various organizations continue to mull over ethical norms to govern AI.[29] However, most early proponents of such norms have moved from the formation stage to the implementation stage. Private companies are currently applying the principles to which they unilaterally subscribed. After having issued one of the first documents on ethical norms,[30] the IEEE is now developing concrete technical standards to be applied by developers to specific applications of AI.[31] ISO, another professional organization, is currently setting such standards as well.[32] Domestic courts and authorities are adjudicating the first cases on AI.[33]

At this point, it is worth pausing for a moment. The current section sketched a process in which multiple actors shaped and formed ethical norms on AI and are now implementing them. (As Section IV will explain, states have not been absent from this process.) This section could now go on to distil the essence of the ethical norms. This would make sense as the ethics remain

understand accountability, keep records, understand the law. 2. Align with user values, inter alia by bringing in policy makers and academics; 3. Keep it explainable, i.e., allow for user questions and make AI reviewable; 4. Minimize bias and promote inclusion. 5. Protect users' data rights, adhere to national and international rights laws.

[27] AI Now, 'AI Now 2017 Report' (*AI Now Institute*, 2017) https://ainowinstitute.org/AI_Now_2017_Report.pdf, recommendation no 10: 'Ethical codes [. . .] should be accompanied by strong oversight and accountability mechanisms.' (p 2); see also AI Now, 'AI Now 2018 Report' (*AI Now Institute*, 2018) https://ainowinstitute.org/AI_Now_2018_Report.pdf, recommendation no 3: 'The AI industry urgently needs new approaches to governance.' (p 4).

[28] Partnership on AI, 'Tenets' www.partnershiponai.org/tenets/ (hereafter Partnership on AI, 'Tenets'), in summary: 1. Benefit and empower as many people as possible; 2. Educate and listen, inform; 3. Be committed to open research and dialogue on the ethical, social, economic, and legal implications of AI; 4. Research and development need to be actively engaged with, and accountable to, stakeholders; 5. Engage with, and have representation of, stakeholders in the business community; 6. Maximize benefits and address challenges by: protecting privacy and security; understanding and respecting interests of all parties impacted; ensuring that the AI community remains socially responsible, sensitive and engaged; ensuring that AI is robust, reliable, trustworthy, and secure; opposing AI that would violate international conventions and human rights; and promoting safeguards and technology that do no harm; 7. Be understandable and interpretable for people for purposes of explaining the technology; 8. Strive for a culture of cooperation, trust, and openness among AI scientists and engineers.

[29] See, for instance, Pontifical Academy for Life, Microsoft, IBM, FAO and Ministry of Innovation (Italian Government), 'Rome Call for AI Ethics' (*Rome Call*, 28 February 2020) www.romecall.org.

[30] IEEE, 'Ethically Aligned Design' (n 10).

[31] See the IEEE P7000 standards series, e.g. IEEE SA, IEEE P7000 - Draft Model Process for Addressing Ethical Concerns During System Design (*IEEE*, 30 June 2016) https://standards.ieee.org/project/7000.html; The IEEE considers standard setting with regard to AI unprecedented: 'This is the first series of standards in the history of the IEEE Standards Association that explicitly focuses on societal and ethical issues associated with a certain field of technology'; IEEE, 'Ethically Aligned Design' (n 10) 283; for the type of standard that is necessary, see D Danks, AJ London, 'Regulating Autonomous Systems: Beyond Standards', (2017) 32 *IEEE Intelligent Systems* 88.

[32] See ISO, 'Standards by ISO / IEC JTC 1 / SC 42. Artificial Intelligence' www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0.

[33] See UK High Court, R (Bridges) v CCSWP and SSHD [2019] EWHC 2341 (Admin); UK Hight Court, R (Bridges) v CCSWP and SSHD [2020] EWCA Civ 1058; Tribunal Administratif de Marseille, La Quadrature du Net, No. 1901249 (27 Nov. 2020); Swedish Data Protection Authority, 'Supervision pursuant to the General Data Protection Regulation (EU) 2016/679 – facial recognition used to monitor attendance of students' (DI-2019-2221, 20 August 2019) <imy.se/globalassets/dokument/beslut/facial-recognition-used-to-monitor-the-attendance-of-students.pdf>; a number of non-governmental organisations are bringing an action against Clearview AI Inc., which sells facial recognition software, for violation of data protection law, see https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe. A global inventory listing incidents involving AI that have taken place so far includes more than 600 entries to date: AIAAIC repository: https://docs.google.com/spreadsheets/d/1Bn55B4xz21-_Rgdr8BBb2lton_4rzLGxFADMlVW0PYI/edit#gid=888071280; compare with AI Incident Database, 'All Incident Reports' (7 June 2021) https://incidentdatabase.ai/, which is run by the Partnership on AI and includes 100 incidents.

unconsolidated and fuzzy. But much important work has already been done in this direction.[34] In fact, for present purposes, no further efforts are necessary because, while norms remain vague, they have now begun to merge into domestic law. However, the diffusion of ethical norms is far from being a linear and straightforward process with clear causes. Instead, it is multidirectional, multivariate, gradual, and open-ended, with plenty of back and forth. Hence, the next section, as it looks at norm diffusion from the incoming end, in other words, from the perspectives of states and domestic law, is best read as a continuation of the present section. The developments outlined have also occurred in parallel to those in municipal law, which are the topic of the next section.

III. DIFFUSION OF ETHICAL NORMS INTO DOMESTIC LAW: THE NEW REGULATION OF THE EUROPEAN UNION ON AI

A relevant sign of diffusion into domestic law is states' first engagement with ethics and AI. For some states, including China, France, Germany, and the United States, such engagement began relatively early with the adoption of AI strategies[35] in which ethical norms figured more or less prominently. The French president, for instance, stated a commitment to establish an ethics framework.[36] China, in its strategy, formulated the aim to '[d]evelop laws, regulations, and ethical norms that promote the development of AI'.[37] Germany's strategy was to task a commission to come up with recommendations concerning ethics.[38] The US strategy, meanwhile, was largely silent on ethics.[39]

[34] A Jobin, M Ienca, and E Vayena, 'The Global Landscape of AI Ethics Guidelines' (2019) 1 *Nature Machine Intelligence* (2019) 389–399; J Fjeld and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (*Berkman Klein Center for Internet & Society*, 2020) http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420.

[35] State Council of the People's Republic of China, 'A Next Generation Artificial Intelligence Development Plan' (*New America*, 20 July 2017) www.newamerica.org/documents/1959/translation-fulltext-8.1.17.pdf (hereafter China, 'AI Development Plan'); President of the French Republic, 'The President of the French Republic Presented His Vision and Strategy to Make France a Leader in AI at the Collège de France on 29 March 2018' (*AI for Humanity*, 2018) www.aiforhumanity.fr/en/ (hereafter French Republic, 'Strategy to Make France a Leader in AI'); Federal Government of Germany, 'Artificial Intelligence Strategy' (*The Federal Government*, November 2018) www.ki-strategie-deutschland.de/home.html?file=files/downloads/Nationale_KI-Strategie_engl.pdf (hereafter Germany, 'AI Strategy'); US President, 'Executive Order on Maintaining American Leadership in Artificial Intelligence' (2019) E.O. 13859 of Feb 11, 2019, 84 FR 3967 (hereafter US President, 'Executive Order on Leadership in AI'). According to T Dutton, 'An Overview of National AI Strategies' (*Medium*, 28 June 2018) https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd which contains a useful list of national AI strategies, Canada was the first state to put forward such a national strategy in the year 2017. Yet it remains unclear what exactly constitutes a 'strategy'. In any case, the documents published by the Obama Administration in 2016 (see n 38) already contained many elements of a 'strategy'.

[36] French Republic, 'Strategy to Make France a Leader in AI' (n 35) third commitment.

[37] China, 'AI Development Plan' (n 35) Section V 1; the text accompanying this aim is more concrete. It recommends addressing traceability and accountability; to launch research on AI behaviour science and ethics; and 'establish an ethical and moral multi-level judgment structure and human-computer collaboration ethical framework'. China is also committed to 'actively participate in global governance of AI, strengthen the study of major international common problems such as robot alienation and safety supervision, deepen international cooperation on AI laws and regulations, international rules and so on, and jointly cope with global challenges'.

[38] Germany, 'AI Strategy' (n 35) 4, 37, 38. The data ethics commission ('Datenethikkommission') in response published its report in October 2019: Datenethikkommission, 'Gutachten' (*BMI*, October 2019) www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=4. The report deals comprehensively on 240 pages with 'digitization', not just AI, and includes 75 recommendations to move forward. An economic assessment of the proposals in the report would be necessary though. The report seems quite 'big' on regulation.

[39] The US strategy merely stated as one of five guiding principles: 'The United States must foster public trust and confidence in AI technologies and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people.' (US President, 'Executive Order on Leadership in AI' (n 35) section 1(d); compare with National Science and Technology Council, 'Preparing for the Future of

Some state legislative organs also addressed the ethics of AI early on, most notably, the comprehensive report published by the United Kingdom House of Lords in 2018.[40] It, among other things, recommended elaborating an AI code to provide ethical guidance and a 'basis for statutory regulation, if and when this is determined to be necessary'.[41] The UK report also suggested five ethical principles as a basis for further work.[42] In a similar vein, the Villani report, which had preceded the French presidential strategy, identified five ethical imperatives.[43]

In the EU, a report drafted within the European Parliament in 2016 drew attention to the need to examine ethics further.[44] It dealt with robotics because AI was not yet a priority and included a code of rudimentary ethical principles to be observed by researchers. In 2017, the European Parliament adopted the report as a resolution,[45] putting pressure on the Commission to propose legislation.[46] In 2018, the Commission published a strategy on AI with a threefold

Artificial Intelligence' (*The White House, President Barack Obama*, October 2016) https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf which had been published before and addressed transparency, fairness, and efficacy of systems in recommendations nos 16 and 17 and ethics in education curricula in recommendation no 20, and National Science and Technology Council, 'The National Artificial Intelligence Research and Development Strategic Plan', (*The White House, President Barack Obama*, October 2016) https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf, which was published on the same day as Preparing for the Future of Artificial Intelligence, p. 3: 'understand and address the ethical, legal, and societal implications of AI' is a research priority according to strategy no. 3. See also the webpage of the US government on AI which has recently gone live: www.ai.gov/.

[40] House of Lords (Select Committee on Artificial Intelligence), 'AI in the UK: Ready, Willing and Able?' (*UK Parliament*, 16 April 2018) https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (hereafter House of Lords, 'AI in the UK').

[41] House of Lords, 'AI in the UK' (n 40) para 420.

[42] House of Lords, 'AI in the UK' (n 40) para 417, in brief: 1. Development of AI for common good and humanity; 2. Intelligibility and fairness; 3. Use of AI should not diminish data rights or privacy; 4. Individuals' right to be educated to flourish mentally, emotionally and economically alongside AI; 5. The autonomous power to hurt, destroy, or deceive human beings should never be vested in AI. In the United Kingdom, further work also addressed the use of facial recognition technology: Biometrics and Forensics Ethics Group (BFEG UK government), 'Interim Report of BFEG Facial Recognition Working Group' (*OGL*, February 2019) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/781745/Facial_Recognition_Briefing_BFEG_February_2019.pdf. According to this report, facial recognition: 1. Is only permissible when in public interest; 2. Justifiable only if effective; 3. Should not involve or exhibit bias; 4. Should be deployed in even-handed ways: for example, not target certain events only (impartiality); 5. Should be a last resort: No other less invasive alternative, minimizing interference with lawful behaviour (necessity). Also, 6. Benefits must be proportionate to loss of liberty and privacy; 7. Humans must be impartial, accountable, oversighted, esp. when constructing watch lists; and 8. Public consultation and rationale are necessary for trust. Finally, 9. Could resources be used better elsewhere?

[43] C Villani, 'For a Meaningful Artificial Intelligence – Towards a French and European Strategy' (*AI for Humanity*, March 2018) www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf 113–114; in summary: 1. transparency and auditability; 2. Rights and freedoms need to be adapted in order to forestall potential abuse; 3. Responsibility; 4. Creation of a diverse and inclusive social forum for discussion; 5. Politicization of the issues linked to technology. Compare with D Dawson and others, 'Artificial Intelligence – Australia's Ethics Framework, A Discussion Paper' (2019) https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf 6, which, in a nutshell, proposed the following ethics guidelines: 1. Generate net benefits; 2. Civilian systems should do no harm; 3. Regulatory and legal compliance; 4. Protection of privacy; 5. Fairness: no unfair discrimination, particular attention to be given to training data; 6. Transparency and explainability; 7. Contestability; 8. Accountability, even if harm was unintended.

[44] Draft Report with recommendations to the Commission on Civil Law Rules on Robotics, 2015/2103(INL), 23 May 2016; the report was marked by an alarmist undertone.

[45] Resolution with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)), European Parliament, P8_TA (2017)0051, 16 February 2018.

[46] Ibid, para 65.

aim, one of which was to ensure 'an appropriate legal and ethical framework'.[47] The Commission consequently mandated a group of experts who suggested guidelines for 'trustworthy' AI one year later.[48] These guidelines explicitly drew on work previously done within the institutions.[49] The guidelines refrained from interfering with the *lex lata*,[50] including the General Data Protection Regulation[51].

In 2019, following the guidelines for trustworthy AI, the Commission published a White Paper on AI[52], laying the foundation for the legislative proposal to be tabled a year later. The White Paper, which attracted much attention,[53] recommended a horizontal approach to AI with general principles included in a single legislative act applicable to any kind of AI, thus rejecting the alternative of adapting existing (or adopting several new) sectorial acts. The White Paper

[47] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial Intelligence for Europe, European Commission, 25 April 2018, section 1 toward the end.

[48] High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (8 April 2019) www.ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf (hereafter: 'Ethics Guidelines for Trustworthy AI'). The Guidelines distinguish between foundations of trustworthy AI which include four ethical principles, namely 1. Respect for human autonomy, 2. Prevention of harm, 3. Fairness, 4. Explicability (12 *et seq*) and seven requirements for their realization, namely 1. Human agency and oversight, 2. Technical robustness and safety, 3. Privacy and data governance, 4. Transparency, 5. Diversity, non-discrimination, fairness, 6. Societal and environmental well-being and 7. Accountability.

[49] Notably European Group on Ethics in Science and New Technologies (EGE), 'Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems' (9 March 2018) https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382>. Another initiative within the wider sphere of the EU worked in parallel with the Commission's High-Level Expert Group and published a set of principles: L Floridi and others, 'AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 *Minds and Machines* 689.

[50] See Ethics Guidelines for Trustworthy AI (n 48) 6: 'The Guidelines do not explicitly deal with the first component of Trustworthy AI (lawful AI), but instead aim to offer guidance on fostering and securing the second and third components (ethical and robust AI).' And 10: 'Understood as legally enforceable rights, fundamental rights therefore fall under the first component of Trustworthy AI (lawful AI), which safeguards compliance with the law. Understood as the rights of everyone, rooted in the inherent moral status of human beings, they also underpin the second component of Trustworthy AI (ethical AI), dealing with ethical norms that are not necessarily legally binding yet crucial to ensure trustworthiness.'

[51] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1 (GDPR). The General Data Protection Regulation, in Article 22 regulates automated decision making and therefore one aspect of AI; however, the effectiveness of the Article is limited by the scope of Regulation as well as loopholes in paragraph 2. Article 22 is entitled 'Automated Individual Decision-Making, Including Profiling' and reads as follows: '1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. 2. Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent. 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision; 4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.' For an international legal perspective on the General Data Protection Regulation, see the Symposium on: 'The GDPR in International Law' (6 January 2020) AJIL Unbound 114.

[52] European Commission, *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, European Commission* (White Paper, COM(2020) 65 final, 2020) (hereafter White Paper on AI).

[53] The public consultation on the White Paper on AI (n 52) attracted a wide range of comments, see e.g. Google, 'Consultation on the White Paper on AI – a European Approach' (*Google*, 28 May 2020) www.blog.google/documents/77/Googles_submission_to_EC_AI_consultation_1.pdf.

suggested regulating AI based on risk: the higher the risk of an AI application, the more regulation was necessary.[54]

On 21 April 2021, based on the White Paper, the Commission presented a Proposal for a regulation on AI[55]. The Commission's Proposal marks a crucial moment, for it represents the first formal step – globally, it seems – in a process that will ultimately lead to binding domestic legislation on AI. It is a sign of the absorption of ethical norms on AI by domestic law – in other words, of norm diffusion. While the risk-based regulatory approach adopted from the White Paper was by and large absent in the ethics documents discussed in the previous section, many of the substantive obligations in the proposed regulation reflect the same ethical norms.

The Commission proposed distinguishing three categories of AI, namely: certain 'practices' of AI that the proposed regulation prohibits; high-risk AI, which it regulates in-depth; and low-risk AI required to be flagged.[56] While the prohibition against using AI in specific ways (banned 'practices')[57] attracts much attention, practically, the regulation of high-risk AI will be more relevant. Annexes II and III to the proposed regulation determine whether an AI qualifies as high-risk.[58] The proposed regulation imposes a series of duties on those who place such high-risk AI on the market.[59]

The regulatory focus on risky AI has the consequence, on the flip side, that not all AI is subject to the same degree of regulation. Indeed, the vast majority of AI is subject merely to the duty to ensure some degree of transparency. However, an AI that now appears to qualify as low-risk

[54] White Paper on AI (n 52) 17: an application of AI should be considered high-risk, when it is situated in a sensitive domain, e.g. health care, and presents a concrete risk.

[55] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, European Commission, COM (2021) 206 final, 21 April 2021, in the following: the Proposal or the proposed regulation.

[56] See Article 52 of the proposed regulation which states a relatively light transparency obligation with regard to AI not presenting high risks ('certain AI systems', according to Article 52).

[57] The regulation proposes to ban the practice of AI: a) to materially distort a person's behaviour (a draft leaked earlier had called this 'manipulation'); b) to exploit the vulnerabilities of a specific group of persons ('targeting' of vulnerable groups, according to the leaked draft); c) social scoring by the public authorities, and d) for live remote biometric identification in public places (see article 5(1)(a)–(d) of the proposed regulation). The regulation does not preclude the development of AI, even if it could eventually be used in ways the regulation prohibits. A pathway is required in the case of letters a and b: the practices are only prohibited if they are at least likely to cause a person physical or psychological harm. The ban of biometric identification according to letter d is subject to a public security exception pursuant to Article 5(2).

[58] The definition of AI in annex I appears to be in accordance with how the term is understood in the computer sciences (compare S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (3rd ed., 2014), but it is a broad definition that lawyers may read differently than computer scientists and the elements added in Article 3(1) of the proposed regulation distort it to some degree. Annex II lists legislative acts of the Union; if an act listed applies (e.g., in case of medical devices or toys), any AI used in this context is to be considered high-risk. Annex III relies on domains in conjunction with concrete, intended uses. It lists the following domains: remote biometric identification systems (if not banned by article 5), critical infrastructure, educational institutions, employment, essential public and private services, law enforcement and criminal law, management of migration, asylum, and border control, as well as assistance of judicial authorities. Specific uses listed under these domains are caught as high-risk AI. For instance, AI is considered high-risk when it is intended to be used for predictive policing (use) in law enforcement (domain). The Commission, jointly with the Parliament and the Council, is delegated the power to add further uses within the existing domains, which, in turn, could only be added to by means of a full legislative amendment; the Commission's power is subject to an assessment of potential harm (see Articles 7 and 73 of the proposed regulation).

[59] Mostly the 'provider' will be the person who puts an AI on the market, according to Article 16 of the proposed regulation; sometimes it is the importer, the distributor or another third party, according to Articles 26–28; Article 3(2) defines a provider as 'a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge'.

under the proposed regulation could become high-risk after a minor change in use intention. Hence, given the versatility of AI, the duties applicable to high-risk AI have to be factored in even in the development of AI in low-risk domains. One example is an image recognition algorithm that *per se* qualifies as low-risk under the regulation. However, if it were later used for facial recognition, the more onerous duties concerning high-risk AI would become applicable. Such development must be anticipated at an early stage to ensure compliance with the regulation throughout the life cycle of AI. Hence, regulatory spill-over from high-risk into low-risk domains of AI is likely. Consequently, the proposed regulation exerts a broader compliance pull than one might expect at first glance, given the specific, narrow focus of the regulation on high-risk AI.

Categorization aside, the substantive duties imposed on those who put high-risk AI on the market are most interesting from the perspective of ethical norm diffusion. The proposed regulation includes four bundles of obligations.

The first bundle concerns *data* and is laid down in Article 10 of the proposed regulation. When AI is trained with data (though not only then[60]), Article 10 of the proposed regulation requires 'appropriate data governance and management practices', in particular concerning design choices; data collection; data preparation; assumptions concerning that which data measures and represents; assessment of availability, quantity, and suitability of data; 'examination in view of possible bias'; and identification of gaps and shortcomings. In addition, the data itself must be relevant, representative, free of errors, and complete. It must also have 'appropriate statistical properties' regarding the persons on whom the AI is used. And it must take into account the 'geographical, behavioural or functional setting' in which the AI will be used.

The duties laid down in Article 10 on data mirror existing ethical norms, notably the imperative to avoid bias. The IEEE's Charter discussed the issue of data bias.[61] In an early set of principles addressed to professionals, avoidance of bias featured prominently; it also recommended keeping a description of data provenance.[62] The Montreal Declaration recommended avoiding discrimination,[63] while the Toronto Declaration on human rights and machine learning had bias and discrimination squarely in view.[64] Likewise, some of the ethical norms the private sector had adopted addressed bias.[65] However, the ethical norms discussed in Section II generally refrained from addressing data and its governance as

[60] Article 10(6) of the proposed regulation transposes some of the requirements applicable to trained AI to AI that has not been trained.

[61] IEEE, 'Ethically Aligned Design' (n 10) 188, recommending careful assessment of bias and integration of potentially disadvantaged groups in the process; Future of Life Institute, 'Asilomar AI Principles' (n 11) did not yet address bias explicitly.

[62] USACM, 'Algorithmic Transparency' (n 12), principle no 1: '1. Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.' Principle no 5 addressed 'data provenance'. Compare Japanese Society for AI, 'Guidelines' (n 12) principle no 5 with a slightly broader scope.

[63] Montreal Declaration for AI (n 22) principle no 6.1: 'AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on – among other things – social, sexual, ethnic, cultural, or religious differences.' See also principle no 7 concerning diversity; there are some data governance requirements in principle no 8 on prudence.

[64] Toronto Declaration (n 22) for instance, no 16. Not all documents laying down ethics principles discuss bias; OpenAI Charter (n 25) for instance, leaves bias aside and focuses on the safety of general AI.

[65] By way of example, Sage, 'The Ethics of Code' (n 26) principle no 1; Google, 'AI Principles' (n 26) principle no 2; IBM, 'Ethics for AI' (n 26) discusses fairness, including avoidance of bias, as one of five ethics principles (34–35); it also includes recommendations on how to handle data: 'Your AI may be susceptible to different types of bias based on the type of data it ingests. Monitor training and results in order to quickly respond to issues. Test early and often.' Partnership on AI, Tenets (n 28) on the other hand, only generically refers to human rights (see tenet no 6.e).

comprehensively as Article 10 of the proposed regulation. Instead, the ethical norms directly focused on avoidance of bias and discrimination.

The second bundle of obligations concerns *transparency* and is contained in Article 13 of the proposed regulation. The critical duty of Article 13 requires providers to 'enable users to interpret [the] output' of high-risk AI and 'use it appropriately'[66]. The article further stipulates that providers have to furnish information that is 'concise, complete, correct and clear'[67], in particular regarding the 'characteristics, capabilities and limitations of performance' of a high-risk AI system.[68] These duties specifically relate to any known or foreseeable circumstance, including foreseeable misuse, which 'may lead to risks to health and safety or fundamental rights', and to performance on persons.[69]

Transparency is an equally important desideratum of ethical norms, though it is sometimes addressed in terms of explainability or explicability. The IEEE's Charter[70] and the Asilomar principles[71] emphasized transparency to different degrees. Other guidelines encourage the production of explanations[72] or appropriate and sufficient information,[73] or call for extensive transparency, justifiability, and intelligibility.[74] These references make it evident that ethical norms, though they are heterogeneous and vague, are in the process of being absorbed by EU law (norm diffusion).

The third bundle of obligations is contained in Article 15 of the proposed regulation. It requires high-risk AI to have an 'appropriate level' of *accuracy*, *robustness*, and *cybersecurity*.[75] Article 15 refrains from adding much detail but states that the AI must be resilient to deleterious environmental influences or nefarious third parties' attempts to game it.[76]

As with the first and second bundles, the aspects of high-risk AI addressed by Article 15 can be traced back to various ethical norms. The high-level principles of effectiveness and awareness of misuse in the IEEE's Charter covered similar aspects.[77] The Asilomar principles addressed 'safety', but in a rather generic fashion.[78] Other principles emphasized both the need for safety in all things related to AI and the importance of preventing misuse.[79] Others focused on prudence, which more or less includes the aspects covered by Article 15.[80] Parts of the private sector also committed themselves to safe AI.[81]

[66] Article 13(1) of the proposed regulation.
[67] Article 13(2) of the proposed regulation.
[68] Article 13(3b) of the proposed regulation.
[69] Article 13(3b)(iii and iv) of the proposed regulation.
[70] IEEE, 'Ethically Aligned Design' (n 10) 11; transparency implies that the basis of a decision of an AI should 'always be discoverable'.
[71] Asilomar AI Principles (n 11) according to principle no 7, it must be possible to ascertain why an AI caused harm; according to principle no 8, any involvement in judicial decision making should be explainable and auditable.
[72] USACM, 'Algorithmic Transparency' (n 12) principle no 4.
[73] Japanese Society for AI, 'Guidelines' (n 12) principle no 5 (addressing security).
[74] Montreal Declaration for AI (n 22) principle no 5, with 10 sub-principles addressing various aspects of transparency. See also The Toronto Declaration (n 22) which includes strong transparency obligations for states (para 32) and weaker obligations for the private sector (para 51).
[75] Article 15(1) of the proposed regulation.
[76] Article 15(3 and 4) of the proposed regulation.
[77] IEEE, 'Ethically Aligned Design' (n 10) 11, principles nos 4 and 7.
[78] Asilomar AI Principles (n 11) principle no 6.
[79] Japanese Society for AI, 'Guidelines' (n 12) principles nos 5 and 7.
[80] 'Montreal Declaration for AI (n 22) principle no 8; The Toronto Declaration (n 22) has a strong focus on non-discrimination and human rights; it does not address the topics covered by Article 15 of the proposed regulation directly. Open AI Charter (n 25) stated a commitment to undertake the research to make AI safe in the long term.
[81] E.g. Google, 'AI Principles' (n 26) principle no 3: 'Be built and tested for safety'; IBM, 'Ethics for AI' (n 26) 42–45, addressed certain aspects of safety and misuse under 'user data rights'. See also Partnership on AI, 'Tenets' (n 28) tenet

The fourth bundle contains obligations of a *procedural or managerial nature*. The proposed regulation places confidence in procedure to cope with the high risks of AI. The trust in procedure goes so far that substantive issues are addressed procedurally only. One such example is one of the cardinal obligations of the proposed regulation, namely the duty to manage risks according to Article 9. Article 9 obliges providers to maintain a comprehensive risk management system throughout the life cycle of high-risk AI. It aims at reducing the risks posed by the AI so that the risks are 'judged acceptable', even under conditions of foreseeable misuse.[82] The means to reduce the risks are design, development, testing, mitigation and control measures, and provision of information to users. Instead of indicating which risks are to be 'judged acceptable', Article 9 trusts that risk reduction will result from a series of diligently executed, proper steps. However, procedural rules are not substantive rules. In and of themselves, they do not contain substantive guidance. In essence, Article 9 entrusts providers with the central 'judgment' of what is 'acceptable'. Providers are granted liberty, while their obligations seem less onerous. At the same time, this liberty imposes a burden on them in that courts might not always validate their 'judgment' of what was 'acceptable' after harm has occurred. Would, for instance, private claims brought against the provider of an enormously beneficial AI be rejected after exceptionally high risks, which the provider managed and judged acceptable, have materialized?

Trust in procedure is also a mainstay of other provisions of the proposed regulation. An assessment of conformity with the proposed regulation has to be undertaken, but, here again, providers carry it out themselves in all but a few cases.[83] Providers have to register high-risk AI in a new EU-wide database.[84] Technical documentation and logs must be kept.[85] Human oversight is required – a notion that has a procedural connotation.[86] The regulation does not require substantive 'human control' as discussed within CCW for autonomous weapons systems.[87] Discrimination is not directly prohibited, but procedural transparency is supposed to contribute to preventing bias.[88] Such transparency may render high-risk AI interpretable, but a substantive right to explicable AI is missing.[89]

The procedural and managerial obligations in the fourth bundle cannot easily be traced back to ethical norms. This is because of their procedural nature. Ethical norms are, in essence, substantive norms. Procedural obligations are geared towards implementation, yet implementation is not the standard domain of ethics (except for applied ethics which is yet to reach AI[90]). Hence, while certain aspects of the fourth bundle mirror ethical norms, for example, the requirement to keep logs,[91] none of them has called for a comprehensive risk management system.

no 6.d: 'Ensuring that AI research and technology is robust, reliable, trustworthy, and operates within secure constraints.'

[82] Article 9(4) of the proposed regulation.

[83] Articles 19 and 43 of the proposed regulation.

[84] Article 60(2) of the proposed regulation.

[85] Articles 11–12 of the proposed regulation.

[86] Human oversight can be either built into AI or measures can be merely identified so that users can appropriately implement them, according to Article 14(3) of the proposed regulation. Oversight should enable users to understand and monitor AI, interpret its output, decide not to use it, intervene in its operation, and prevent automation bias (Article 14(4)).

[87] See Eleven Guiding Principles on Lethal Autonomous Weapons (n 19); note that 'meaningful human control' is not mentioned as a requirement for autonomous weapons systems in these guiding principles.

[88] See the discussion of bias above.

[89] See the discussion of transparency above.

[90] But see Trusilo and Burri, 'Ethical AI' (n 16).

[91] See, for instance, USACM, 'Algorithmic Transparency' (n 12) principle no 6: 'Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.' (Emphasis removed.)

Overall, the proposed regulation offers compelling evidence of norm diffusion, at least to the extent that the regulation reflects ethical norms on AI. It addresses the three most pressing concerns related to AI of the machine learning type, namely bias due to input data, opacity that hampers predictability and explainability, and vulnerability to misuse (gaming, etc.).[92] In addressing these concerns, the proposed regulation remains relatively lean. It notably refrains from taking on broader concerns with which modern AI is often conflated, namely dominant market power,[93] highly stylized concepts,[94] and the general effects of technology.[95]

However, the proposed regulation does not fully address the main concerns concerning AI, namely bias and opacity, head-on. It brings to bear a gentle, procedural approach on AI by addressing bias indirectly through data governance and transparency and remedying opacity through interpretability. It entrusts providers with the management of the risks posed by AI and with the judgement of what is tolerable. Providers consequently bear soft duties. In relying on soft duties, the regulation extends the life of ethical norms and continues their approach of indulgence. It thus incorporates the character of ethical norms that lack the commitment of hard law.

On the one hand, it may be unexpected that ethical norms live on to a certain extent, given that the new law on AI is laid down in a directly applicable, binding Union regulation. On the other hand, this is not all that surprising because a horizontal legislative act that regulates all kinds of AI in one go is necessarily less specific on substance than several sectorial acts addressing individual applications. (Though the adoption of several sectorial acts would have had other disadvantages.) Yet, this approach of the proposed regulation begs the question of whether it can serve as a basis for individual, private rights: will natural persons, market competitors, etc. be able to sue providers of high-risk AI for violation of the procedural, managerial obligations incumbent on them under the regulation?[96]

IV. INTERNATIONAL LAW SIDELINED

It is not the case that international law has ignored the rise of AI, while ethics filled the void and laid down the norms. International law – especially the soft type – and ethical principles overlap and are not always easily distinguishable. Yet, even international soft law has been lagging behind considerably. It took until late spring 2019 for the Organization for Economic Co-Operation and Development (OECD) to adopt a resolution spelling out five highly abstract principles on AI.[97] While the principles address opacity (under transparency and explainability)

[92] The risk of a responsibility gap is not addressed by the proposed regulation, but by a revision of the relevant legislation on liability, see p 5 of the Explanatory Memorandum to the proposed regulation.

[93] See A Ezrachi and ME Stucke, *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy* (2016).

[94] Bostrom, 'Superintelligence' (n 8); J Dawes, 'Speculative Human Rights: Artificial Intelligence and the Future of the Human' (2020) 42 *Human Rights Quarterly* 573.

[95] For a broader perspective on AI, see K Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (2021).

[96] Note the broad geographical scope of the proposed regulation. It applies when providers bring AI into circulation in the Union, but also when output produced outside of the Union is used in it (see Article 2(1)(a) and (c) of the proposed regulation). The substantive scope of the proposed regulation is not universal, though, for it, for instance, largely excludes weapons and cars (see Article 2(2) and (3) of the proposed regulation).

[97] OECD Recommendation OECD/LEGAL/0449 of 22 May 2019 of the Council on Artificial Intelligence (hereafter OECD, 'Recommendation on AI'; the five principles are the following: 1. Inclusive growth, sustainable development and well-being; 2. Human-centred values and fairness; 3. Transparency and explainability; 4. Robustness, security and safety; 5. Accountability. Another five implementing recommendations advise specifically States to: invest in AI research and development; foster a digital ecosystem; shape the policy environment for AI, including by way of

and robustness (including security and safety), they ignore the risk of bias. Instead, they only generically refer to values and fairness. When the OECD was adopting its non-binding resolution, the European Commission's White Paper[98] was already in the making. As the White Paper, the OECD Resolution recommended a risk-based approach.[99] Additionally, the OECD hosts a recent political initiative, the Global Partnership on Artificial Intelligence,[100] which has produced a procedural report.[101]

Regional organizations have been more alert to AI than universal organizations. Certain sub-entities of the Council of Europe notably examined AI in their specific purview. In late 2018, a commission within the Council of Europe adopted a set of principles governing AI in the judicial system;[102] in the Council of Europe's data protection convention framework, certain principles focussing on data protection and human rights were approved in early 2019.[103] On the highest level of the Council of Europe, the Committee of Ministers recently adopted a recommendation,[104] which discussed AI ('algorithmic

experimentation; build human capacity and prepare for labour market transformation; and cooperate internationally, namely on principles, knowledge sharing, initiatives, technical standards, and metrics; see also S Voeneky, 'Key Elements of Responsible Artificial Intelligence – Disruptive Technologies, Dynamic Law' (2020) 1 *Ordnung der Wissenschaft* 9, 16.

[98] White Paper on AI (n 52).

[99] OECD'Recommendation on AI' (n 97) point 1.4.c.

[100] OECD, 'OECD to Host Secretariat of New Global Partnership on Artificial Intelligence' (OECD, 15 June 2020) https://www.oecd.org/newsroom/oecd-to-host-secretariat-of-new-global-partnership-on-artificial-intelligence.htm; the idea of this initiative may be to counterweigh China in AI: J Delcker, 'Wary of China, the West Closes Ranks to Set Rules for Artificial Intelligence' (*Politico*, 7 June 2021) www.politico.eu/article/artificial-intelligence-wary-of-china-the-west-closes-ranks-to-set-rules/. The OECD initiative is not to be confused with the Partnership on Artificial Intelligence, see Partnership on AI, 'Tenets' (n 28).

[101] The Global Partnership on Artificial Intelligence, Responsible Development, Use and Governance of AI, Working Group Report (GPAI Summit Montreal, November 2020) www.gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2020.pdf.

[102] European Commission for the Efficiency of Justice (CEPEJ), 'European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment' (*Council of Europe*, 3-4 December 2018) https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c. In sum, it suggested the following guidelines: 1. Ensure compatibility with human rights; 2. Prevent discrimination; 3. Ensure quality and security; 4. Ensure transparency, impartiality, and fairness: make AI accessible, understandable, and auditable; 5. Ensure user control.

[103] Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, 'Guidelines on Artificial Intelligence and Data Protection (Council of Europe Convention 108)' (25 January 2019) T-PD(2019)01. The guidelines distinguish between general principles (i), principles addressed to developers (ii), and principles addressed to legislators and policy makers (iii). In summary, the principles are the following: i) 1. Respect human rights and dignity; 2. Respect the principles of Convention 108+: lawfulness, fairness, purpose specification, proportionality of data processing, privacy-by-design and by default, responsibility and demonstration compliance (accountability), transparency, data security and risk management; 3. Avoid and mitigate potential risks; 4. Consider functioning of democracy and social/ethical values; 5. Respect the rights of data subjects; 6. Allow control by data subjects over data processing and related effects on individuals and society. ii) 1. Value-oriented design; 2. Assess, precautionary approach; 3. Human rights by design, avoid bias; 4. Assess data, use synthetic data; 5. Risk of decontextualised data and algorithms; 6. Independent committee of experts; 7. Participatory risk assessment; 8. Right not to be subject solely to automated decision making; 9. Safeguard user freedom of choice to foster trust, provide feasible alternatives to AI; 10. Vigilance during entire life-cycle; 11. Inform, right to obtain information; 12. Right to object. iii) 1. Accountability, risk assessment, certification to enhance trust; 2. In procurement: transparency, impact assessment, vigilance; 3. Sufficient resources for supervisors. 4. Preserve autonomy of human intervention; 5. Consultation of supervisory authorities; 6. Various supervisors (data, consumer protection, competition) should cooperate; 7. Independence of committee of experts in ii.6; 8. Inform and involve individuals; 9. Ensure literacy. See also Consultative Committee of the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, 'Guidelines on Facial Recognition (Convention 108)' T-PD(2020)03rev4.

[104] Recommendation CM/Rec(2020)1 of 8 April 2020 of the Committee of Ministers to Member States on the human rights impacts of algorithmic systems, Council of Europe Committee of Ministers, (hereafter 'Recommendation on

systems',[105] as it calls it) in depth from a human rights perspective. The recommendation drew the distinction between high-risk and low-risk AI that the proposed Union regulation also adopted.[106] It, in large parts, mirrors the European Union's approach developed in the White Paper and the proposed regulation. This is not surprising given the significant overlap in the two organizations' membership.

On the universal level, processes to address AI have moved at a slower pace. The United Nations Educational, Scientific and Cultural Organization is only now discussing a resolution addressing values, principles, and fields of action on a highly abstract level.[107] The United Nations published a High-Level Report in 2019,[108] but it dealt with digital technology and its governance from a general perspective. Hence, the values it lists[109] and the recommendations it makes[110] appear exceedingly abstract from an AI point of view. The three models of governance suggested in the report, however, break new ground.[111]

In a nutshell, most of the international law on AI arrives too late. Domestic implementation of ethical norms is already in full swing. Legislative acts, such as the proposed regulation of the EU, are already being adopted. Court and administrative cases are being decided. Meanwhile,

the human rights impacts'). The recommendation is a detailed text that first addresses states and then private actors. After elaborating on scope and context (part A paras 1–15, discussing, for example, synthetic data [para 6], the fusion of the stages of development and implementation of AI [para 7], the presence of both private and public aspect in many algorithmic systems [para 12], and a precautionary approach [para 15]), it lists obligations of states in part B, including data management (para 2), testing (paras 3.3–5), transparency and remedies (para 4), and precautionary measures (para 5, including standards and oversight). These obligations are then tailored to the situation of private actors on the basis of the due diligence approach applicable to business. The obligations in this part are less stringent; see, for instance, the duty to prevent discrimination in para C.1.4.

[105] Recommendation on the human rights impacts (n 104) para A.2.

[106] Recommendation on the human rights impacts (n 104) para A.11.

[107] See UNESCO, 'Draft text of the Recommendation on the Ethics of Artificial Intelligence' SHS/IGM-AIETHICS/2021/APR/4 (*UNESCO Digital Library*, 31 March 2021) https://unesdoc.unesco.org/ark:/48223/pf0000376713; see also UNESCO, 'Artificial Intelligence for Sustainable Development: Challenges and Opportunities for UNESCO's Science and Engineering Programmes' SC/PCB/WP/2019/AI (*UNESCO Digital Library*, August 2019); see F Molnár-Gábor, Die Herausforderung der medizinischen Entwicklung für das internationale soft law am Beispiel der Totalsequenzierung des menschlichen Genoms, (2012) 72 *Zeitschrift für ausländisches öffentliches Recht und Völkerrecht* 695, for the role of soft law created by UNESCO.

[108] UN High Level Panel on Digital Cooperation, 'The Age of Digital Interdependence: Report of the UN Secretary-General's High-Level Panel on Digital Cooperation' (*UN*, June 2019) (hereafter 'The Age of Digital Interdependence').

[109] The Age of Digital Interdependence (n 108) 7: Inclusiveness, respect, human-centredness, human flourishing, transparency, collaboration, accessibility, sustainability, and harmony. That 'values' are relative in AI becomes evident from the key governance principles the Report lays down in Section VI. The principles, each of which is explained in one sentence, are the following: Consensus-oriented; Polycentric; Customised; Subsidiarity; Accessible; Inclusive; Agile; Clarity in roles and responsibility; Accountable; Resilient; Open; Innovative; Tech-neutral; Equitable outcomes. Further key functions are added: Leadership, Deliberation; Ensuring inclusivity; Evidence and data; Norms and policy making; Implementation; Coordination; Partnerships; Support and Capacity development; Conflict resolution and crisis management. This long list that appears like the result of a brainstorming begs the question of the difference between the 'values' of the Report on page 7 and the 'principles' ('functions') on page 39 and how they were categorized.

[110] The Age of Digital Interdependence (n 108) 29–32; the recommendations include: 1B: Creation of a platform for sharing digital public goods; 1C: Full inclusion for women and marginalized groups; 2: Establishment of help desks; 3A: Finding out how to apply existing human rights instruments in the digital age; 3B: Calling on social media to work with governments; 3C: Autonomous systems: explainable and accountable, no life and death decisions, non-bias; 4: Development of a Global Commitment on Digital Trust and Security; 5A: By 2020, create a Global Commitment for Digital Cooperation; welcoming a UN Technology envoy.

[111] The Age of Digital Interdependence (n 108) 23–26: The three governance models that are proposed are the following: i) a beefed-up version of the existing Internet governance forum; ii) a distributed, multi-stakeholder network architecture, which to some extent resembles the status quo; and iii) an architecture that is more government driven, while it focuses on the idea of 'digital commons'.

standardization organizations are enacting the technical – and not-so-technical – details. Still, the international law on AI, all of which is soft (and hence not always distinguishable from 'ethical norms'), is far from being useless. The Council of Europe's recommendation on algorithmic systems[112] added texture and granularity to the existing ethical norms. Instruments that may eventually be adopted on the universal level may spread norms on AI across the global south and shave off some of the Western edges the norms (and AI itself) currently still carry.[113]

However, the impact of the ethical norms on AI is more substantial than international legal theory suggests. The ethical norms were consolidated outside of the traditional venues of international law. By now, they are diffusing into domestic law. International law is a bystander in this process. Even if the formation of formally binding international law on AI were attempted at some point,[114] a substantial treaty would be hard to achieve as domestic legislatures would have locked in legislation by then. A treaty could only re-enact a consensus established elsewhere, in other words, in ethical norms and domestic law, which would reduce its compliance pull.

V. CONCLUSION AND OUTLOOK

This chapter explained how ethical norms on AI came into being and are now absorbed by domestic law. The European Union's new proposal for a regulation on AI illustrated this process of 'bottom-to-bottom' norm diffusion. While soft international law contributed to forming ethical norms, it neither created them nor formed their basis in a formal, strict legal sense.

This chapter by no means suggests that law always functions or is created in the way illustrated above. Undoubtedly, international law is mainly formed top-down through classical sources. In this case, it also exercises compliance pull. However, in domains such as AI, where private actors – including multinational companies and transnational or domestic non-governmental organizations – freely shape the landscape, a transnational process of law creation takes place. States in such cases tend to realize that 'their values' are at stake when it is already too late. Hence, states and their traditional way of making international law are sidelined. However, it is not ill will that drives the process of norm diffusion described in this chapter. States are not deliberately pushed out of the picture. Instead, ethical norms arise from the need of private companies and individuals for normative guidance – and international law is notoriously slow to deliver it. When international law finally delivers, it does not set the benchmark but only re-traces ethical norms. However, it does at least serve to make them more durable, if not inalterable.

The discussion about AI in international law has so far been about the international law that should, in a broad sense, govern AI. Answers were sought to how bias, opacity, robustness, etc., of AI could be addressed and remedied through law. However, a different dimension of international law has been left out of the picture so far. Except for the narrow discussion about autonomous weapons systems within CCW, international lawyers have mainly neglected what

[112] Recommendation on the human rights impacts (n 104).

[113] See the useful mapping of AI in emerging economies: 'Global South Map of Emerging Areas of Artificial Intelligence' (*K4A*, 9 June 2021) www.k4all.org/project/aiecosystem/; Knowledge for All, a foundation, conducts useful projects on development and AI, see www.k4all.org/project/?type=international-development.

[114] The Council of Europe is currently deliberating on whether to draft a treaty on AI: Feasibility Study, Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI), CAHAI(2020)23.

AI means for international law itself and the concepts at its core.[115] Therefore, the next step to be taken has to include a re-assessment of central notions of international law in the light of AI. The notions of territoriality/jurisdiction, due diligence duties concerning private actors, control that is central to responsibility of all types, and precaution should consequently be re-assessed and recalibrated accordingly.

[115] A further dimension relates to the use of AI for international lawyers, see A Deeks, 'High-Tech International Law' (2020) 88(3) *George Washington Law Review* 574–653; M Scherer, 'Artificial Intelligence and Legal Decision-Making: The Wide Open? — A Study Examining International Arbitration' (2019) 36 *Journal of International Arbitration* (5) 539–574; for data analysis and international law, see W Alschner, 'The Computational Analysis of International Law' in R Deplano and N Tsagourias (eds), *Research Methods in International Law: A Handbook* (2021) 204–228.

8

Fostering the Common Good

An Adaptive Approach Regulating High-Risk AI-Driven Products and Services

*Thorsten Schmidt and Silja Voeneky**

I. INTRODUCTION

The risks based on AI-driven systems, products, and services are human-made, and we as humans are responsible if a certain risk materialises and damage is caused. This is one of the main reasons why States and the international community as a whole should prioritise governing and responsibly regulating these technologies, at least if high-risks are plausibly linked to AI-based products or services.[1] As the development of new AI-driven systems, products, and services is based on the need of private actors to introduce new products and methods in order to survive as part of the current economic system,[2] the core and aim of the governance and regulative scheme should not hinder responsible innovation by private actors, but minimize risks as far as possible for the common good, and prevent violations of individual rights and values – especially of legally binding human rights. At least the protection of human rights that are part of customary international law is a core obligation for every State[3] and is not dependent on the respective constitutional framework or on the answer as to which specific international human rights treaty binds a certain State.[4]

* Thorsten Schmidt and Silja Voeneky are grateful for the support and enriching discussions at Freiburg Institute for Advanced Studies (FRIAS). Thorsten Schmidt wants to thank Ernst Eberlein, and Silja Voeneky all members of the interdisciplinary FRIAS Research Group *Responsible AI* for valuable insights. Besides, Voeneky's research has been financed as part of the interdisciplinary research project *AI Trust* by the Baden-Württemberg Stiftung (since 2020). Earlier versions of parts of Sections II-IV of this Chapter have been published in S Voeneky, 'Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks' in S Voeneky and G Neuman (eds), *Human Rights, Democracy and Legitimacy in a World of Disorder* (2018) 139 *et seq.* and S Voeneky, 'Key Elements of Responsible Artificial Intelligence: Disruptive Technologies, Dynamic Law' (2020) 1 *OdW* 9 *et seq.*

[1] This approach is part of the concept of 'Responsible AI'. In the following, we concentrate on a regulative approach for high-risk AI-driven products; we nevertheless include – for a regulation *mutatis mutandis* – AI-based high-risk services.

[2] J Beckert and R Bronk, 'An Introduction to Uncertain Futures' in J Beckert and R Bronk (eds), *Uncertain Futures: Imaginaries, Narratives, and Calculation in the Economy* (2018), who link this to the capitalist system, only, which seems like a too narrow approach.

[3] Human rights treaties do *not* oblige non-state actors, such as companies; however, States are obliged to respect, protect, and fulfill human rights and the due diligence framework can be applied in the field of human rights protection; *cf.* M Monnheimer, *Due Diligence Obligations in International Human Rights Law* (2021) 13 *et seq.*, 49 *et seq.*, 204 *et seq.* With regard to a human-rights based duty of States to avoid existential and catastrophic risks that are based on research and technological development, *cf.* S Voeneky, 'Human Rights and Legitimate Governance of Existential and Global Catastrophic Risks' in S Voeneky and G Neuman (eds), *Human Rights, Democracy and Legitimacy in a World of Disorder* (2018) 139, 151 *et seq.* (hereafter Voeneky, 'Human Rights and Legitimate Governance').

[4] It is still disputed, however, whether there is an obligation for States to regulate extraterritorial corporate conduct, *cf.* M Monnheimer, *Due Diligence Obligations in International Human Rights Law* (2021) 307 *et seq.* For a positive answer Voeneky, 'Human Rights and Legitimate Governance' (n 3) 155 *et seq.*

In this chapter, we want to spell out core elements of a regulatory regime for high-risk AI-based products and such services that avoid the shortcomings of regimes relying primarily on preventive permit procedures (or similar preventive regulation) and that avoid, at the same time, the drawbacks of liability-centred approaches. In recent times both regulative approaches failed in different areas to be a solid basis for fostering justified values, such as the right to life and bodily integrity, and protecting common goods, such as the environment. This chapter will show that – similar to regulating risks that stem from the banking system – risks based on AI products and services can be diminished if the companies developing and selling the products or services have to pay a proportionate amount of money into a fund as a financial guarantee after developing the product or service but before market entry. We argue that it is reasonable for a society, a State, and also the international community to adopt rules that oblige companies to pay such financial guarantees to supplement preventive regulative approaches and liability norms. We will specify what amount of money has to be paid based on the *ex-ante* evaluation of risks linked to the high-risk AI product or AI-based service that can be seen as proportionate, in order to minimize risks, but fostering responsible innovation and the common good. Lastly, we will analyse what kind of accompanying regulation is necessary to implement the approach proposed by us. *Inter alia*, we suggest that a group of independent experts should serve as an expert commission to assess the risks of AI-based products and services and collect data on the effects of the AI-driven technology in real-world settings.

Even though the EU Commission has recently drafted a regulation on AI (hereafter: Draft EU AIA),[5] it is not the purpose of this chapter to analyze this proposal in detail. Rather, we intend to spell out a new approach that could be implemented in various regulatory systems in order to close regulatory gaps and overcome disadvantages of other approaches. We argue that our proposed version of an '*adaptive*' *regulation* is compatible with different legal systems and constitutional frameworks. Our proposal could further be used as a blueprint for an international treaty or international soft law[6] declaration that can be implemented by every State, especially States with companies that are main actors in developing AI-driven products and services.

The term AI is broadly defined for this chapter, covering the most recent AI systems based on complex statistical models of the world and the method of machine learning, especially self-learning systems. It also includes systems of classical AI, namely, AI systems based on software already programmed with basic physical concepts (preprogrammed reasoning),[7] as a symbolic-reasoning engine.[8] AI in its various forms is a multi-purpose tool or general purpose technology

[5] Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on AI (Artificial Intelligence Act) and amending certain Union Legislative Acts' COM(2021) 206 final.

[6] See Section II.

[7] For a broad definition see as well the Draft EU AIA; according to this AI system "means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions influencing the environments they interact with." Article 3(1) and Annex I Draft EU AIA reads: "(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; (c) Statistical approaches, Bayesian estimation, search and optimization methods."

[8] *Cf.* recently M Bhatt, J Suchan, and S Vardarajan, 'Commonsense Visual Sensemaking for Autonomous Driving: On Generalised Neurosymbolic Online Abduction Integrating Vision and Semantics' (2021) 299 *Artificial Intelligence Journal* https://doi.org/10.1016/j.artint.2021.103522. Here we are concerned with the concept of 'object permanence', in other words, the idea that "discrete objects continue to exist over time, that they have spatial relationships with one another (such as in-front-of and behind)", the understanding that objects, such as cars, continue to exist even if they disappear behind an obstacle; see also 'Is a Self-Driving Car Smarter Than a Seven-Month-Old?' *The Economist* (2021) www.economist.com/science-and-technology/is-it-smarter-than-a-seven-month-old/21804141.

and a rapidly evolving, innovative key element of many new and possibly disruptive technologies applied in many different areas.[9] A recent achievement, for instance, is the merger of biological research and AI, demonstrated by the use of an AI-driven (deep-learning) programme that a company can use to determine the 3D shapes of proteins.[10] Moreover, applications of AI products and AI-based services exist not only in the areas of speech recognition and robotics but also in the areas of medicine, finance, and (semi-)autonomous cars, ships, planes, or drones. AI-driven products and AI-driven services already currently shape areas as distinct as art or weapons development.

It is evident that potential risks accompany the use of AI-driven products and services and that the question of how to minimize these risks without impeding the benefits of such products and services poses great challenges for modern societies, States, and the international community. These risks can be caused by actors that are not linked to the company producing the AI system as these actors might misuse an AI-driven technology.[11] But damages can also originate from the unpredictability of adverse outcomes (so-called off-target effects[12]), even if the AI-driven system is used for its originally intended purpose. Damage might also arise because of a malfunction, false or unclear input data, flawed programming, etc.[13] Furthermore, in some areas, AI services or products will enhance or create new systemic risks. For example, in financial applications[14] based on deep learning,[15] AI serves as a cost-saving and highly efficient tool and is applied on an increasingly larger scale. The uncertainty of how the AI system reacts in an unforeseen and untested scenario, however, creates new risks, while the large-scale implementation of new algorithms or the improvement of existing ones additionally amplifies already existing risks. At the same time, algorithms have the potential to destabilize the whole financial system,[16] possibly leading to dramatic losses depending on the riskiness and the implementation of the relevant AI-driven system.

[9] S Russel and P Novig, *Artificial Intelligence: A Modern Approach* (3rd ed., 2016), 1. S Voeneky, 'Key Elements of Responsible Artificial Intelligence – Disruptive Technologies, Dynamic Law' (2020) 1 *OdW* 9, 10–11 with further references (hereafter Voeneky, 'Key Elements of Responsible Artificial Intelligence') https://ordnungderwissenschaft.de/wp-content/uploads/2020/03/2_2020_voeneky.pdf; I Rahwan and others, 'Machine Behaviour' (2019) *Nature* 568, 477–486 (2019) www.nature.com/articles/s41586-019-1138-y; for the various fields of application *cf.* also W Wendel, 'The Promise and Limitations of Artificial Intelligence in the Practice of Law' (2019) 72 *Oklahoma Law Review* 21, 21–24, https://digitalcommons.law.ou.edu/olr/vol72/iss1/3/.

[10] This might be a tool to solve the so-called protein folding problem, *cf.* E Callaway, '"It Will Change Everything": DeepMind's AI Makes Gigantic Leap in Solving Protein Structures' (2020) 588 *Nature* 203 www.nature.com/articles/d41586-020-03348-4.

[11] M Brundage and others, 'The Malicious Use of Artificial Intelligence' (*Malicious AI Report*, 2018) https://maliciousaireport.com/ 17.

[12] For this notion in the area of biotechnology, *cf.* XH Zhang and others, 'Off-Target Effects in CRISPR/Cas9-Mediated Genome Engineering' (2015) 4 *Molecular Therapy: Nucleic Acids* https://doi.org/10.1038/mtna.2015.37; WA Reh, *Enhancing Gene Targeting in Mammalian Cells by the Transient Down-Regulation of DNA Repair Pathways* (2010) 22.

[13] *Cf.* C Wendehorst in this volume, Chapter 12.

[14] Such as high-frequency trading, deep calibration, deep hedging and risk management. High-frequency trading means the automated trading of securities characterized by extremely high speeds and high turnover rates; deep calibration means the fitting of a model to observable data of derivatives (calibration) by deep neural networks and deep hedging means the derivation of hedging strategies by the use of deep neural networks. For details on the topic of AI and finance, *cf.* M Paul, Chapter 21, in this volume.

[15] To list a few examples of this rapidly growing field, *cf.* J Sirignano and R Cont, 'Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning' (2019) 19(9) *Quantitative Finance* 1449–1459; H Buehler and others, 'Deep Hedging' (2019) 19(8) *Quantitative Finance* 1271–1291; B Horvath, A Muguruza, and M Tomas, 'Deep Learning Volatility: A Deep Neural Network Perspective on Pricing and Calibration in (Rough) Volatility Models' (2021) 21(1) *Quantitative Finance* 11–27.

[16] J Danielsson, R Macrae, and A Uthemann, 'Artificial Intelligence and Systemic Risk' (*Systemic Risk Centre*, 24 October 2019) www.systemicrisk.ac.uk/publications/special-papers/artificial-intelligence-and-systemic-risk.

Even more, we should not ignore the risk posed by the development of so-called superhuman AI: Because recent machine learning tools like reinforcement learning can improve themselves without human interaction and rule-based programming,[17] it seems to be possible for an AI system – as argued by some scholars – to create an improved AI system which opens the door to produce some kind of artificial Superintelligence or superhuman AI (or 'the Singularity').[18] Superhuman AI might even pose a global catastrophic or existential risk to humanity.[19] Even if some call this a science-fiction scenario, other experts predict that AI of superhuman intelligence will happen by 2050.[20] It is argued, as well, that an intelligence explosion might lead to dynamically unstable systems and it becomes increasingly easy for smarter systems to make themselves smarter[21] that finally, there can be a point beyond which it is impossible for us to make reliable predictions.[22] In the context of uncertainty and 'uncertain futures',[23] it is possible that predictions fail and risks arise from these developments faster than expected or in an unexpected fashion.[24] From this, we deduce that superhuman AI can be seen as a low probability, high impact scenario.[25] Because of the high impact, States and the international community should not ignore the risks of superhuman AI when drafting rules concerning AI governance.

II. KEY NOTIONS AND CONCEPTS

Before spelling out in more detail lacunae and drawbacks of the current specific regulation governing AI-based products and services, there is a need to define key notions and concepts relevant for this chapter, especially the notions of regulation, governance, and risk.

When speaking about governance and regulation, it is important to differentiate between legally binding rules on the one hand at the national, European, and international level, and non-binding soft law on the other hand. Only the former are part of the law and regulation *strictu sensu*.

[17] *See* Y LeCun and others, 'Deep Learning' (2015) 521 *Nature* 436–444 www.nature.com/nature/journal/v521/n7553/full/nature14539.html.

[18] The term 'the Singularity' was coined in 1993 by the computer scientist Vernon Vinge; he argued that "[w]ithin thirty years, we will have the technological means to create superhuman intelligence," and he concluded: "I think it's fair to call this event a singularity ('the Singularity' (. . .))." *See* V Vinge, 'The Coming Technological Singularity: How to Survive in the Post-Human Era' in GA Landis (ed), *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace* (1993) 11, 12.

[19] See also in this volume J Tallinn and R Ngo, Chapter 2. *Cf.* as well S Hawking, 'Will Artificial Intelligence Outsmart Us?' in S Hawking (ed), *Brief Answers to the Big Questions* (2018), 181; S Russel and P Novig, *Artificial Intelligence: A Modern Approach* (3rd ed., 2016) 1036 *et seq.*; S Bringsjord and NS Govindarajulu, 'Artificial Intelligence' in EN Zalta (ed), *The Stanford Encyclopedia of Philosophy* (2020) https://plato.stanford.edu/entries/artificial-intelligence/ 9; A Eden and others, *Singularity Hypotheses: A Scientific and Philosophical Assessment* (2013); A Al-Imam, MA Motyka, and MZ Jędrzejko, 'Conflicting Opinions in Connection with Digital Superintelligence' (2020) 9(2) *IAES IJ-AI* 336–348; N Bostrom, *Superintelligence* (2014) esp. 75 (hereafter N Bostrom, *Superintelligence*); K Grace and others, 'When Will AI Exceed Human Performance? Evidence from AI Experts' (2018) 62 *Journal of Artificial Intelligence Research* 729–754 https://doi.org/10.1613/jair.1.11222.

[20] See *e.g.*, R Kurzweil, *The Singularity Is Near* (2005) 127; for more forecasts, see Bostrom, *Superintelligence* (n 14) 19–21.

[21] E Yudkowsky, 'Artificial Intelligence as a Positive and Negative Factor in Global Risk' in N Bostrom, MM Ćirković (eds), *Global Catastrophic Risks* (2011) 341.

[22] M Tegmark, 'Will There Be a Singularity within Our Lifetime?' in J Brockman (ed), *What Should We Be Worried About?* (2014) 30, 32.

[23] See for this J Beckert and R Bronk, 'An Introduction to Uncertain Futures' in J Beckert and R Bronk (eds), *Uncertain Futures: Imaginaries, Narratives, and Calculation in the Economy* (2018) 1–38, 2 who argue that 'actors in capitalist systems face an open and indeterminate future'.

[24] As argued in E Yudkowsky, 'There's No Fire Alarm for Artificial General Intelligence' (*Machine Intelligence Research Institute*, 13 October 2017) https://intelligence.org/2017/10/13/fire-alarm/.

[25] Voeneky, 'Human Rights and Legitimate Governance' (n 3) 150.

The term international soft law is understood in this chapter to include rules that cannot be attributed to a formal legal source of public international law and that are, hence, not directly legally binding. However, as rules of international soft law have been agreed upon by subjects of international law (i.e. States, International Organizations (IO)) that could, in principle, create international law[26] these rules possess a specific normative force and can be seen as relevant in guiding the future conduct of States, as they promised not to violate them.[27] Therefore, rules of international soft law are part of top down rulemaking, (i.e. regulation), and must not be confused with (bottom up) private rulemaking by corporations, including the many AI related codes of conduct, as for example, the Google AI Principles.[28]

In the following, regulation means only top down law making by States at the national, and European level or by States and IOs at the international level. It will not encompass rulemaking by private actors that is sometimes seen as an element of so-called self-regulation. However, in the following, the notion of governance will include rules that are part of top-down lawmaking (e.g. international treaties and soft law) and rules, codes, and guidelines by private actors.[29]

Another key notion for the adaptive governance framework we are proposing is the notion of risk. There are different meanings of 'risk' and in public international law, there is no commonly accepted definition of the notion, it is unclear how and whether a 'risk' is different from a 'threat', a 'danger', or a 'hazard'.[30] For the sake of this chapter, we will rely on the following broad definition, according to which a risk is an unwanted event that may or may not occur,[31] that is, an unwanted hypothetical future event. This definition includes situations of uncertainty, where no probabilities can be assigned for the occurrence of damage.[32] A global catastrophic risk

[26] For a similar definition, see D Thürer, 'Soft Law' in R Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (2012) volume 9 271 para 8.

[27] On the advantages and disadvantages of 'standards' compared to 'regulation' see J Tate and G Banda, 'Proportionate and Adaptive Governance of Innovative Technologies: The Role of Regulations, Guidelines, Standards' (*BSI*, 2016) www.bsigroup.com/localfiles/en-gb/bis/innovate%20uk%20and%20emerging%20technologies/summary%20report%20-%20adaptive%20governance%20-%20web.pdf 14 (hereafter Tate and Banda, 'Proportionate and Adaptive Governance').

[28] AI Google, 'Artificial Intelligence at Google: Our Principles' https://ai.google/principles/.

[29] It is beyond the scope of this chapter to discuss bottom up rules drafted by corporations or NGOs in the area of AI.

[30] See G Wilson, 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law' (2013) 31 *Va Envtl LJ* 307, 310. Sometimes there is no differentiation made between threat, hazard, and risk, *see* OECD Recommendation OECD/LEGAL/040 of 6 May 2014 of the Council on the Governance of Critical Risks www.oecd.org/gov/risk/Critical-Risks-Recommendation.pdf. For details see Voeneky, 'Human Rights and Legitimate Governance' (n 3) 140 *et seq*.

[31] See SO Hansson, 'Risk' in EN Zalta (ed), *Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/entries/risk/. In a quantitative sense, risk can be defined through risk measures (be it relying on probabilities or without doing so). Typical examples specify risk as to the probability of an unwanted event that may or may not occur (value-at-risk); or as the expectation of an unwanted event that may or may not occur (expected shortfall). The expectation of a loss materialized by the unwanted event is the product of its size in several scenarios with the probability of these scenarios and thus specifies an average loss given the unwanted event. Many variants of risk measures exist, see for example AJ McNeil, R Frey, and P Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition* (2015). Adaptive schemes rely on conditional probabilities whose theory goes back to T Bayes, 'An Essay Towards Solving a Problem in the Doctrine of Chances' (1764) 53 *Phil Transactions* 370. In the area of international law, the International Law Commission (ILC) stated that the 'risk of causing significant transboundary harm' refers to the combined effect of the probability of occurrence of an accident and the magnitude of its injurious impact, *see* ILC, 'Draft Articles on Prevention of Transboundary Harm from Hazardous Activities' (2001) 2(2) *YB Int'l L Comm* 152.

[32] For a different, narrower notion of risk, excluding situations of uncertainty ('uncertainty versus risk'), see CR Sunstein, *Risk and Reason: Safety, Law and the Environment* (2002)129; CR Sunstein, *Worst-Case Scenarios* (2007)146–147; RA Posner, *Catastrophe* (2004) 171. A judge of the International Court of Justice (ICJ), however, included 'uncertain risks' into the notion of risks, see ICJ, *Pulp Mills on the River of Uruguay (Argentina v Uruguay)*, Sep Op of Judge Cançado Trindade [2010] ICJ Rep 135, 159, 162; for a similar approach (risk as 'unknown dangers') see J Peel, *Science and Risk Regulation in International Law* (2010) 1.

shall be defined as a hypothetical future event that has the potential to cause the death of a large number of human beings or/and to cause the destruction of a major part of the earth; and an existential risk can be defined as a hypothetical future event that has the potential to cause the extinction of human beings on earth.[33]

When linking AI-driven products and services to high-risks, we understand high-risks as those that have the potential to cause major damages for protected individual values and rights (as life and bodily integrity) or common goods (as the environment or the financial stability of a State).

The question of which AI systems, products, or services constitute such high-risk systems is discussed in great detail. The EU Commission has presented a proposal in 2021 as the core element of its Draft EU AIA regulating high-risk AI systems.[34] According to the Draft EU AIA, high-risk AI systems shall include, in particular, human-rights sensitive AI systems, such as AI systems intended to be used for the biometric identification and categorization of natural persons, AI systems intended to be used for the recruitment or selection of natural persons, AI systems intended to be used to evaluate the creditworthiness of natural persons, AI systems intended to be used by law enforcement authorities as polygraphs, and AI systems concerning the area of access to, and enjoyment of, essential private services, public services, and benefits as well as the area of administration of justice and democratic processes, thereby potentially affecting the rule of law in a State (Annex III Draft EU AIA). Nevertheless, it is open for debate whether high-risk AI products and services might include as well, because of the possibility to cause major damages, (semi-)autonomous cars, planes, drones, and ships, and certain AI-driven medical products (such as brain–computer-interfaces, mentioned below) or AI-driven financial trading systems.[35]

Additionally, autonomous weapons clearly fall under the notion of high-risk AI products. However, AI-driven autonomous weapon systems constitute a special case due to the highly controversial ethical implications and the international laws of war (*ius in bello*) governing their development and use.[36]

Another particular case of high-risk AI systems are AI systems that are developed in order to be part of or constitute superhuman AI – some even classify these AI systems as global catastrophic risks or existential risks.

[33] For slightly different definitions, see N Bostrom, 'Superintelligence' (n 12) 115 (stating that '[a]n existential risk is one that threatens to cause the extinction of Earth-originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development'; and N Bostrom and MM Ćirković, 'Introduction' in N Bostrom and MM Ćirković (eds), *Global Catastrophic Risks* (2008) arguing that a *global catastrophic* risk is a hypothetical future event that has the potential 'to inflict serious damage to human well-being on a global scale'.

[34] *Cf.* n 5.

[35] For a definition of high-risk AI products by the European Parliament (EP), *cf.* EP Resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics, and related technologies (2020/2021(INL)), para 14: 'Considers, in that regard, that artificial intelligence, robotics and related technologies should be considered high-risk when their development, deployment and use entail a significant risk of causing injury or harm to individuals or society, in breach of fundamental rights and safety rules as laid down in Union law; considers that, for the purposes of assessing whether AI technologies entail such a risk, the sector where they are developed, deployed or used, their specific use or purpose and the severity of the injury or harm that can be expected to occur should be taken into account; the first and second criteria, namely the sector and the specific use or purpose, should be considered cumulatively.' www.europarl.europa.eu/doceo/document/TA-9-2020-10-20_EN.html#sdocta9.

[36] Autonomous weapons are expressly outside the scope of the Draft EU AIA, *cf.* Article 2(3).

III. DRAWBACKS OF CURRENT REGULATORY APPROACHES OF HIGH-RISK AI PRODUCTS AND SERVICES

To answer the most pressing regulative and governance questions concerning AI-driven high-risk products and such services, this chapter introduces an approach for responsible governance that shall supplement existing rules and regulations in different States. The approach, spelled out below in more detail, is neither dependent on, nor linked to, a specific legal system or constitutional framework of a specific State. It can be introduced and implemented in different legal cultures and States, notwithstanding the legal basis or the predominantly applied regulatory approach. This seems particularly important as AI-driven high-risk products and such services are already being used and will be used to an even greater extent on different continents in the near future, and yet the existing regulatory approaches differ.

For the sake of this chapter, the following simplifying picture might illustrate relevant general differences: some States rely primarily on a preventive approach and lay down permit procedures or similar preventive procedures to regulate emerging products and technologies;[37] they even sometimes include the rather risk-averse precautionary principle, as it is the case according to EU law in the area of the EU policy of the environment.[38] The latter intends to oblige States to protect the environment (and arguably other common goods) even in cases of scientific uncertainty.[39] Other States, such as the United States, in many sectors, avoid strict permit procedures altogether or those with high approval thresholds or avoid a strict implementation, and rather rely on liability rules that give the affected party, usually the consumer, the possibility to sue a company and get compensation if a product or service has caused damage.

Both regulative approaches – spelling out a permit or similar preventive procedures, with regard to high-risk products or services in the field of emerging technologies, or liability regimes to compensate consumers and other actors after they have been damaged by using a high-risk product – even if they are combined have major deficits and have to be supplemented. On the one hand, preventive permit procedures are often difficult to implement and might be easy to circumvent, especially in an emerging technology field. This was illustrated in recent years in different fields, including emerging technologies, as by the aircraft MAX 737 incidents[40] or the

[37] The Draft AIA by the EU Commission spells out a preventive approach and does not include any relevant liability rules. However, the Commission has announced the proposal of EU rules to address liability issues related to new technologies, including AI systems in 2022, *cf.* C Wendehorst, Chapter 12, in this volume.

[38] See for the precautionary principle as part of EU law: Article 191(2) Treaty on the Functioning of the European Union, OJ 2016 C202/47 as well as Commission, 'Communication on the Precautionary Principle' COM(2000) 1 final. The precautionary principle (or: approach) is reflected in international law in Principle 15 of the Rio Declaration which holds that: 'In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, *lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures* to prevent environmental degradation.' (Emphasis added), United Nations, 'Rio Declaration on Environment and Development' (UN Conference on Environment and Development, 14 June 1992) UN Doc A/CONF 151/26/Rev 1 Vol I, 3; *cf.* also M Schröder, 'Precautionary Approach/Principle' in R Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (2012) volume 8, 400, paras 1–5. In philosophy, there has been an in-depth analysis and defense of the principle in recent times, *cf.* D Steel, *Philosophy and the Precautionary Principle: Science, Evidence, and Environmental Policy* (2014).

[39] It is also argued that this principle shall be applied in all cases of scientific uncertainty and not only in order to protect the environment, *cf.* C Phoenix and M Treder, 'Applying the Precautionary Principle to Nanotechnology' (*CRN*, January 2004) http://crnano.org/precautionary.htm; N Bostrom, 'Ethical Issues in Advanced Artificial Intelligence' (2003) https://nickbostrom.com/ethics/ai.html 2.

[40] As shown in T Sgobba, 'B-737 MAX and the Crash of the Regulatory System' (2019) 6(4) *Journal of Space Safety Engineering* 299; D Scharper, 'Congressional Inquiry Faults Boeing and FAA Failures for Deadly 737 Max Plane Crashes' *NPR News* (16 September 2020) www.npr.org/2020/09/16/913426448/congressional-inquiry-faults-boeing-and-

motorcar diesel gate[41] cases. If this is the case, damage caused by products after they entered the market cannot be avoided. On the other hand, liability regimes that allow those actors and individuals who suffered damage by a product or service to claim compensation, have the drawback that it is unclear how far they prevent companies from selling unsafe products or services.[42] Companies rather seem to be nudged to balance the (minor and unclear) risk to be sued by a consumer or another actor in the future with the chance to make (major) profits by using a risky technology or selling a risky product or service in the present.

How standard regulatory approaches fail was shown, *inter alia*, by the opiate crisis cases[43] in the United States.[44] Even worse, an accountability gap is broadened if companies can avoid or limit justified compensatory payments in the end via settlements or by declaring bankruptcy.[45]

faa-failures-for-deadly-737-max-plane-cr, key mistakes in the regulatory process were: 'excessive trust on quantitative performance requirements, inadequate risk-based design process, and lack of independent verification by experts.' It is argued that similar failures can happen in many other places, see for example P Johnston and H Rozi, 'The Boeing 737 MAX Saga: Lessons for Software Organizations' (2019) 21(3) *Software Quality Professional* 4.

[41] C Oliver and others, 'Volkswagen Emissions Scandal Exposes EU Regulatory Failures' *Financial Times* (30 September 2015) www.ft.com/content/03cdb23a-6758-11e5-a57f-21b88f7d973f; M Potter, 'EU Seeks More Powers over National Car Regulations after VW Scandal' *Reuters* (27 January 2017) www.reuters.com/article/us-volkswagen-emissions-eu-regulations-idUSKCN0V51IO.

[42] With regard to the disadvantages of the US tort system, MU Scherer, 'Regulating Artificial Intelligence' (2016) 29 *Harvard Journal of Law & Technology* 353, 388, and 391.

[43] The opiate crisis cases in the United States show in an alarming way that insufficient and low threshold regulation that allows to prescribe and sell a high-risk product without reasonable limits cannot be outweighed *ex post* by a liability regime, even if damaged actors claim compensation and sue companies that caused the damage, *cf.* District Court of Cleveland County, *State of Oklahoma, ex rel. Hunter v Purdue Pharma LP*, Case No CJ-2017-816 (2019).

[44] Another example are the actions of oil drilling companies, as the oil drill technology can be seen as a high-risk technology. As part of the the so-called 2010 Deepwater Horizon incident British Petroleum (BP) has caused an enormous marine oil spill. In 2014, US District Court for the Eastern District of Louisiana ruled that BP was guilty of gross negligence and willful misconduct under the US Clean Water Act (CWA). The Court found the company to have acted 'recklessly' (*cf.* US District Court for the Eastern District of Louisiana, *Oil Spill by the Oil Rig 'Deepwater Horizon' in the Gulf of Mexico on April 20, 2010*, Findings of Fact and Conclusion of Law, Phase One Trial, Case 2:19-md-02179-CJB-SS (4 September 2014) 121–122). In another case Royal Dutch Shell (RDS) was sued as its subsidiary in Nigeria had caused massive environmental destruction; the Court of Appeal in The Hague ordered in 2021 that RDS has to pay compensation to residents of the region and begin the purification of contaminated waters (*cf.* Gerechtshof Den Haag, *de Vereniging Milieudefensie v Royal Dutch Shell PLC and Shell Petroleum Development Company of Nigeria LTD/Shell Petroleum Development Company of Nigeria LTD v Friday Alfred Akpan*, 29 January 2021); see E Peltier and C Moses, 'A Victory for Farmers in a David-and-Goliath Environmental Case' *The New York Times* (29 January 2021) www.nytimes.com/2021/01/29/world/europe/shell-nigeria-oil-spills.html.

[45] This, as well, the opioid crisis cases in the United States have shown. *Cf.* J Hoffmann, 'Purdue Pharma Tentatively Settles Thousands of Opioid Cases' *New York Times* (11 September 2019) www.nytimes.com/2019/09/11/health/purdue-pharma-opioids-settlement.html: 'Purdue Pharma (…) would file for bankruptcy under a tentative settlement. Its signature opioid, OxyContin, would be sold by a new company, with the proceeds going to plaintiffs'. In September 2021, a federal bankruptcy judge gave conditional approval to a settlement devoting potentially $10 billion to fighting the opioid crisis but will shield the company's former owners, members of the Sackler family, from any future lawsuits over opioids, see J Hoffmann, 'Purdue Pharma Is Dissolved and Sacklers Pay $4.5 Billion to Settle Opioid Claims' *New York Times* (1 September 2021) www.nytimes.com/2021/09/01/health/purdue-sacklers-opioids-settlement.html. Several US states opposed the deal and planned to appeal against it, *cf.* 'What is the bankruptcy "loophole" used in the Purdue Pharma settlement?' *The Economist* (3 September 2021) www.economist.com/the-economist-explains/2021/09/03/what-is-the-bankruptcy-loophole-used-in-the-purdue-pharma-settlement. See also the Attorney General of Washington's statement of 1 September 2021: "This order lets the Sacklers off the hook by granting them permanent immunity from lawsuits in exchange for a fraction of the profits they made from the opioid epidemic — and sends a message that billionaires operate by a different set of rules than everybody else".

IV. SPECIFIC LACUNAE AND SHORTCOMINGS OF CURRENT AI REGULATION

If we take a closer look at the existing specific regulation and regulatory approaches to AI-driven products and (rarely) services, specific drawbacks become apparent at the national, supranational, and international level. It would be beyond the scope of this chapter to elaborate on this in detail,[46] but some loopholes and shortcomings of AI-specific rules and regulations shall be discussed below.[47]

1. *EU Regulation of AI-Driven Medical Devices*

A first example is the EU Regulation on Medical Devices (MDR),[48] which governs certain AI-driven apps in the health sector and other AI-driven medical devices such as in the area of neurotechnology.[49] The amended MDR was adopted in 2017 and entered into force in 2021.[50] It lays down a so-called scrutiny process[51] for high-risk products (certain class III devices) only, which is a consultation procedure prior to market. It regulates, *inter alia*, AI-driven medical device brain stimulation products, for example, brain–computer-interfaces (BCIs). They are governed by the MDR even if there is no intended medical purpose;[52] thus, the MDR also governs consumer neurotechnology devices.

However, it is a major drawback that AI-driven neurotechnology devices are regulated by the MDR, but this law does not lay down a permit procedure to ensure safety standards and only spells out the less strict scrutiny process. In this aspect, the regulation of AI systems intended for brain stimulation in the EU differs significantly from the regulations governing the development of drugs and vaccines in the EU which lay down rules with significantly higher safety thresholds, including clinical trials and human subjects research.[53] Considering the risks because of the use

[46] For this section see Voeneky, 'Key Elements of Responsible Artificial Intelligence' (n 9) 9 et seq.

[47] This does not include a discussion of AI and data protection regulations. However, the European General Data Protection Regulation (GDPR) aims to protect personal data of natural persons (Article 1(1) GDPR) and applies to the processing of this data even by wholly automated means (Article 2(1) GDPR). See Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, in force since 25 May 2018, OJ 2016 L119/1. The GDPR spells out as well a 'right to explanation' regarding automated decision processes; *cf.* T Wischmeyer, 'Artificial Intelligence and Transparency: Opening the Black Box' in T Wischmeyer and T Rademacher (eds), *Regulating Artificial Intelligence* (2019) 75 and 89: Article 13(2)(f) and 14(2)(g) as well as Article 22 GDPR contain an obligation to inform the consumer about the 'logic involved' as well as 'the significance and the envisaged consequences of such processing for the data subject' but not a comprehensive right to explanation.

[48] Regulation (EU) 2017/745 of the European Parliament and of the Council of 05 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, OJ 2017 L117/1. Besides, AI-based medical devices fall within the scope of high-risk AI systems according to Article 6(1) in conjunction with Annex II (11) Draft EU AIA that explicitly refers to Regulation 2017/745, if such AI systems are safety components of a product or themselves products and subject to third party conformity assessment, *cf.* this Section 3(b).

[49] According to Article 2 MDR 'medical device' '(. . .) means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: (. . .)'. For exemptions see, however, Article 1(6) MDR.

[50] The amended MRD came into force in May 2017, but medical devices are subject to a transition period of three years to meet the new requirements. This transition period was extended until 2021 due to the COVID-19 pandemic, *cf.* Regulation (EU) 2020/561 of the European Parliament and of the Council of 23 April 2020 amending Regulation (EU) 2017/745 on medical devices, as regards the dates of application of certain of its provisions.

[51] *Cf.* Articles 54, 55, and 106(3), Annex IX Section 5.1, and Annex X Section 6 MDR.

[52] Annex XVI: '(. . .) 6. Equipment intended for brain stimulation that apply electrical currents or magnetic or electromagnetic fields that penetrate the cranium to modify neuronal activity in the brain. (. . .)'.

[53] §§ 21 *et seq*. *Arzneimittelgesetz* (AMG, German Medicinal Products Act), BGBl 2005 I 3394; Article 3(1) Regulation (EC) 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down Community procedures

of brain–computer-interfaces to humans and their health and integrity, it is unclear why the regulatory threshold is different from the development and use of drugs. This is even more true if neurotechnology is used as a 'pure' consumer technology by individuals and does not have a particular justification for medical reasons. Besides, there is no regulation of neurotechnology at the international level, and so far, no international treaty obliges the States to minimize or mitigate the risks linked to the use of AI-driven neurotechnology.[54]

2. *National Regulation of Semi-Autonomous Cars*

A second example of sector-specific (top down) regulation for AI-driven products with clear disadvantages that entered already in force are the rules governing semi-autonomous cars in Germany. The relevant German law, the *Straßenverkehrsgesetz*, hereafter Road Traffic Act, was amended in 2017[55] to include new automated AI-based driving systems.[56] From a procedural point of view it is striking that the law-making process was finalized before the federal ethics commission had published its report on this topic.[57] The relevant § 1a (1) Road Traffic Act states that the operation of a car employing a highly or fully automated (this means level 3, but not autonomous (not level 4 and 5))[58] driving function is permissible, provided that the function is used for its intended purpose:

> *Der Betrieb eines Kraftfahrzeugs mittels hoch- oder vollautomatisierter Fahrfunktion ist zulässig, wenn die Funktion bestimmungsgemäß verwendet wird.*[59]

It is striking that the meaning of the notions 'intended purpose' is not laid down by the Road Traffic Act itself or by an executive order but can be defined by the automotive company as a

for the authorization and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency, OJ 2004 L 136/1.

[54] The AI recommendation drafted by the OECD, *cf.* OECD Recommendation, OECD/LEGAL/0449 of 22 May 2019 of the Council on Artificial Intelligence https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 are also insufficient in this respect due to their non-binding soft law character, in more detail Voeneky, 'Key Elements of Responsible Artificial Intelligence', 17 *et seq.* and this Section at 3 a. However, at least some States such as Chile and France are attempting to regulate this area of AI: as part of the Chilean constitutional reform, the current Article 19 of the *Carta Fundamental* is to be supplemented by a second paragraph that protects mental and physical integrity against technical manipulation; *cf.* on the current status of the legislative process: *Cámara dediputadas y diputados*, Boletín No 13827-19 for an English translation of the planned amendment see www.camara.cl/verDoc.aspx?prmID=14151&prmTIPO=INICIATIVA, Anexo 1, p. 14. Furthermore, the implementation of specific 'neurorights' is planned, *cf.* project Boletín No 13828-19. The French bioethics law (*Loi n° 2021-1017 du 2 août 2021 relative à la bioéthique*), which came into force at the beginning of August 2021, allows the use of brain-imaging techniques only for medical and research purposes (Articles 18 and 19), *cf.* www.legifrance.gouv.fr/jorf/id/JORFTEXT000043884384/.

[55] *Straßenverkehrsgesetz* (StVG), *cf.* Article 1 Achtes Gesetz zur Änderung des Straßenverkehrsgesetzes (8. StVGÄndG), BGBl 2017 I 1648.

[56] §§ 1a, 1b and § 63 Road Traffic Act. For an overview of the most relevant international, European, and national rules governing autonomous or automated vehicles, *cf.* E Böning and H Canny , 'Easing the Brakes on Autonomous Driving' (FIP 1/2021) www.jura.uni-freiburg.de/de/institute/ioeffr2/downloads/online-papers/FIP_2021_01_BoeningCanny_AutonomousDriving_Druck.pdf (hereafter Böning and Canny, 'Easing the Brakes').

[57] Germany, Federal Ministry of Transport and Digital Infrastructure, Ethics Commission, 'Automated and Connected Driving' (*BMVI*, June 2017), www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html.

[58] An act regulating fully autonomous cars has been in force since 2021 and has changed the Road Traffic Act, see especially the new §§ 1 d-1g Road Traffic Act. For the draft, *cf.* German Bundestag, 'Entwurf eines Gesetzes zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes (*Gesetz zum autonomen Fahren*)' (Draft Act for Autonomous Driving) (9 March 2021), Drucksache 19/27439 https://dip21.bundestag.de/dip21/btd/19/274/1927439.pdf.

[59] § 1a (1) Road Traffic Act.

private actor producing and selling the cars.[60] Therefore, the Road Traffic Act legitimizes and introduces insofar the private standard-setting by corporations. This provision thus contains an 'opening clause' for self-regulation by private actors but is, as such, too vague.[61] This is an example of a regulatory approach that does not provide sufficient standards in the area of an AI driven product that can be linked to high risks. Hence, it can be argued that the § 1a (1) Road Traffic Act violates the *Rechtsstaatsprinzip*, rule of law, as part of the German Basic Law,[62] which states that legal rules must be clear and understandable for those whom they govern.[63]

3. *General AI Rules and Principles: International Soft Law and the Draft EU AI Regulation*

The question arises whether the lacunae mentioned before at the national and European level in specific areas of AI regulation can be closed by rules of international law (a) and the future regulation at the European level, that is, the 2021 Draft AIA (b).

a. International Regulation? International Soft Law!

So far, there does not exist an international treaty regulating AI systems, products, or services. Nor is such a regulation being negotiated. The aims of the States, having their companies and national interests in mind, are still too divergent. This situation differs from the area of biotechnology, a comparable innovative and as well potentially disruptive technology. Biotechnology is regulated internationally by the the Cartagena Protocol, an international treaty, and this international biotech regulation is based on the rather risk averse precautionary principle.[64] Since more than 170 States are parties to the Cartagena Protocol,[65] one can speak of an almost universal regulation, even if the United States, as a major player, is not a State party and not bound by the Cartagena Protocol. However, even in clear high-risk areas of AI development, such as the development and use of autonomous weapons, an international treaty is still lacking. This contrasts with other areas of high-risk weapons development, such as those of biological weapons.[66]

Nevertheless, as a first step, at least international soft law rules have been agreed upon that spell out the first general principles governing AI systems at the international level. The Organization for Economic Co-operation and Development (OECD) has issued an AI Recommendation in 2019 (hereafter OECD AI Recommendation).[67] Over 50 States have

[60] Böning and Canny, 'Easing the Brakes' (n 56).

[61] This seems true even if the description of the intended purpose and the level of automation shall be 'unambiguous' according to the rationale of the law maker, *cf.* German Bundestag, 'Entwurf eines Gesetzes zur Änderung des Straßenverkehrsgesetzes' (Draft Act for Amending the Road Traffic Act) (2017), Drucksache 18/11300 20 https://dip21.bundestag.de/dip21/btd/18/113/1811300.pdf: '*Die Systembeschreibung des Fahrzeugs muss über die Art der Ausstattung mit automatisierter Fahrfunktion und über den Grad der Automatisierung unmissverständlich Auskunft geben, um den Fahrer über den Rahmen der bestimmungsgemäßen Verwendung zu informieren.*'

[62] Grundgesetz für die Bundesrepublik Deutschland (GG), BGBl 1949 I 1, last rev 29 September 2020, BGBl 2020 I 2048.

[63] B Grzeszick, 'Article 20' in T Maunz und G Dürig (eds), *Grundgesetz-Kommentar* (August 2020), para 99. This is not the case, however, with regard to level 4 and 5 autonomous cars, as the rules enshrined in the 2021 §§ 1 d-1 g Road Traffic Act are more detailed, even including some norms for a a solution of the so-called trolley problem, *cf.* § 1 e para. 2 (no 2).

[64] *Cf.* Section III.

[65] Cartagena Protocol on Biosafety to the Convention on Biological Diversity (adopted 29 January 2000, entered into force 11 September 2003) 2226 UTNS 208.

[66] Convention on the prohibition of the development, production, and stockpiling of bacteriological (biological) and toxin weapons and on their destruction (adopted 10 April 1972, entered into force 26 March 1975) 1015 UNTS 163.

[67] OECD AI Recommendation (n 54).

agreed to adhere to these principles, including States especially relevant for AI research and development, such as the United States, the UK, Japan, and South Korea. The OECD AI Recommendation states and executes five complementary value-based principles:[68] these are inclusive growth, sustainable development, and well-being (IV. 1.1); human-centred values and fairness (IV. 1.2.); transparency and explainability (IV. 1.3.); robustness, security, and safety (IV. 1.4.); and accountability (IV. 1.5.). In addition, AI actors – meaning those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI[69] – should respect the rule for human rights and democratic values (IV. 1.2. lit. a). These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.

However, the wording of the OECD soft law principles is very soft ('should respect'). Even the OECD AI Recommendation on transparency and explainability (IV. 1.3.) has little substance. It states that

> [...] [AI Actors][70] should provide meaningful information, appropriate to the context, and consistent with the state of art: [...]
>
> to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

Assuming that discrimination and unjustified biases are one of the key problems of AI systems,[71] asking for a 'systematic risk management approach' to solve these problems,[72] seems insufficient as a standard of AI actors' due diligence.

Moreover, the OECD AI Recommendation does not mention any legal liability or legal responsibility. AI actors 'should be accountable'. This indicates that these actors should report and provide certain information about what they are doing to ensure 'the proper functioning of AI systems' and 'for the respect of the above principles' (IV. 1.5). This does not imply any legal obligation to achieve these standards or any legal liability if an actor fails to meet the threshold.

Finally, the OECD AI Recommendation does not stress the responsibility of governments to protect human rights in the area of AI. They include only five recommendations to policymakers of States ('adherents', section 2) that shall be implemented in national policies and international cooperation consistent with the above-mentioned principles. These include investing in AI research and development (V. 2.1), fostering a digital ecosystem for AI (V. 2.2), shaping and enabling policy environment for AI (V. 2.3), building human capacity and preparing for labour market transformation (V. 2.4), and international cooperation for trustworthy AI (V. 2.5). Hence, even if an actor aims to rely on the OECD AI Recommendation, it remains unclear what State obligations follow from human rights with regard to the governance of AI.

[68] An AI system is defined as 'a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.' *Cf.* OECD AI Recommendation (n 54).

[69] OECD AI Recommendation (n 54).

[70] AI actors here are 'those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI', see OECD AI Recommendation (n 54).

[71] See Data Ethics Commission, Opinion of the Data Ethics Commission (*BMJV*, 2019), 194 www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3.

[72] 'AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.' *Cf.* IV. 1.4. c) OECD AI Recommendation (n 54).

Besides this, the problem of how to frame the low probability/high risk scenarious (or the low probability/catastrophic or existential risk challenges) linked to the possible development of superhuman AI is not even mentioned in the OECD AI Recommendation.[73]

b. Draft EU AI Regulation

As mentioned above, the draft regulation issued by the European Commission, the Draft EU AIA, proposes harmonized rules on AI systems and spells out the framework for general regulation of AI. It is laying down criteria with regard to requirements for the design and development of high-risk AI systems, not limited to specific sectors. For this, the regulation follows a risk-based regulatory approach – however not based on the precautionary principle – and, at its core, includes a classification of high-risk AI systems, on the one hand, and non-high-risk AI systems, on the other hand. For this, the notion of an AI system is defined in broad terms (Article 3(1) Draft EU AIA).[74] Also, the regulation governs all providers[75] 'placing on the market or putting into service AI systems in the EU' and all users of AI systems in the EU (Article 2, Article 3(2) Draft EU AIA). What kind of AI systems are high-risk AI systems, is laid down in general terms in Articles 6-7 and listed in Annex II and Annex III Draft EU AIA. The Annex III list, mentioned above,[76] can be amended and modified by the EU Commission in the future, which promises that the regulation might not be inflexible regulating the fast-moving field of AI systems as an emerging technology.[77]

The Draft EU AIA aims to limit the possible negative effects of the use of an AI system with regard to the protection of human rights, stressing core human rights as the protection of human dignity, autonomy, and bodily integrity. Therefore, certain 'AI practices' are prohibited according to Article 5 Draft EU AIA, especially if used by State authorities. This includes, but is not limited to, the use of certain AI systems that 'deploy[s] subliminal techniques beyond a person's consciousness' if this is likely to cause harm for a person. The same is true if AI practices cause harm to persons because they exploit the vulnerabilities of a specific group due to their age or disability, or the use of AI systems for law enforcement if this means to use a real-time remote biometric identification system. However, the latter prohibitions are not absolute as exemptions are enshrined in Article 5 Draft EU AIA.

Transparency obligations shall also protect human rights, as there is the need to make it transparent if an AI system is intended to interact with natural persons (Article 52 Draft EU AIA). The same is true with regard to the duty to report 'serious incidents or any malfunctioning (...) which constitutes a breach of obligations under Union law intended to protect fundamental rights' (Article 62 Draft EU AIA).

Apart from these prohibitions and duties, every high-risk AI system must comply with the specific requirements (Article 8 Draft EU AIA). This means that, *inter alia*, risk management systems must be established and maintained (Article 9 Draft EU AIA); training data sets must meet quality criteria (Article 10 Draft EU AIA). Besides, the criteria for the technical

[73] See note 36.

[74] See Section II.

[75] Providers are not limited to private actors but every natural or legal person, including public authorities, agencies, and other bodies, *cf.* Article 3(2).

[76] See Section II.

[77] The European Commission is entitled in Article 7 to add new high-risk systems to Annex III if those systems pose a risk to fundamental rights and safety that is comparable to those systems that are already contained in Annex III. However, this flexibility means that there is only a very loose thread of democratic legitimacy for the future amendments of Annex III. It is beyond the scope of this chapter to discuss this in more detail, but it is unclear whether this disadvantage is sufficiently justified because of the benefit to achieve more flexibility with regard to the regulation of AI systems as a fast-moving technology.

documentation of high-risk AI systems are spelled out in the Draft EU AIA (Article 11 and Annex IV); the operating high-risk AI systems shall be capable of the automatic recording of events and their operation has to be 'sufficiently transparent' (Article 12 and 13 Draft EU AIA). Finally, there must be human oversight (Article 14 Draft EU AIA); the latter could be interpreted as prohibiting the aim to develop and produce superhuman AI.

Another characteristic is that not only developing companies, providers of high-risk AI systems (Article 16 *et seq*. Draft EU AIA), importers and distributors (Articles 26 and 27 Draft EU AIA), but also users are governed by the Draft EU AIA and have obligations. Users encompass companies, as credit institutions, that are using high-risk AI systems (Articles 3(4), together with Articles 28 and 29 Draft EU AIA). Obligations are, for instance, that 'input data is relevant in view of the intended purpose of the high-risk AI system', and the duty to monitor the operation and keep the logs (Article 29 Draft EU AIA).

As the Draft EU AIA includes no relevant liability rules, it is a clear example of a preventive regulatory approach.[78] However, the Draft EU AIA does not establish a permit procedure but only a so-called conformity assessment procedure (Article 48 and Annex V Draft EU AIA), that is either based on internal control (Annex VI Draft EU AIA) or including the involvement of a notified body (Article 19 and 43, Annex VII Draft EU AIA). Notified bodies have to verify the conformity of high-risk AI systems (Article 33 Draft EU AIA). But it is up to the EU Member States to establish such a notifying authority (Article 30 Draft EU AIA) according to the requirements of the Draft EU AIA, and a notified body is allowed to subcontract specific tasks (Article 34 Draft EU AIA). As an oversight, the EU Commission can investigate cases 'where there are reasons to doubt' whether a notified body fulfills the requirements (Article 37 Draft EU AIA).

It has to be mentioned that derogations from the conformity assessment procedure are part of the regulation; derogations exist 'for exceptional reasons of public security or the protection of life and health of persons, environmental protection' and even (*sic!*) 'the protection of key industrial and infrastructure assets' (Article 47 Draft EU AIA).

In the end, many obligations rest on the providers, as for instance the documentation obligations (Article 50 Draft EU AIA), the post-market monitoring (Article 61 Draft EU AIA), or the registration of the system as part of the EU database (Articles 51 and 60 Draft EU AIA). However, if one evaluates how effective an implementation might be, it is striking that the regulation lays down only fines 'up to' a certain amount of money, as 10.000.000–30.000.000 EUR, if the Draft EU AIA is violated and it is up to the EU Member States to decide upon the severity of the penalties. Additionally, administrative fines that could be imposed on Union institutions, agencies, and bodies are much lower ('up to' 250.000 EUR – 500.000 EUR according to Article 72 Draft EU AIA).[79]

It is beyond the scope of this chapter to assess the Draft EU AIA in more detail.[80] Nevertheless, one has to stress that no permit procedure is part of the regulation of high-risk AI systems. This means that this regulation establishes lower thresholds with regard to high-risk AI systems compared, for instance, with the regulation of the development of drugs and vaccines in the EU. It seems doubtful whether the justification provided in the explanatory notes is convincing; it states that a combination with strong ex-post enforcement is an effective and

[78] For this differentiation, *cf.* Section III. For more details *cf.* C Wendehorst, Chapter 12, in this volume.
[79] For enforcement details *cf.* Articles 63 *et seq.*; for penalties *cf.* Article 71.
[80] For details *cf.* T Burri, Chapter 7, in this volume.

reasonable solution, given the early phase of the regulatory intervention and the fact the AI sector is very innovative and expertise for auditing is only now being accumulated.[81]

In the end, without a regulative solution for liability issues, it seems doubtful whether the major risks of high-risk AI systems can be sufficiently mitigated on the basis of the Draft EU AIA. Therefore, another approach shall be proposed by us, one that is compatible with the Draft EU AIA but will complement it to fill in the loopholes.

4. Interim Conclusion

From what has been written above, one can conclude, firstly, that there are loopholes and drawbacks in the regulation of emerging technologies and especially AI systems, although there are rules in place in at least some areas of AI-driven products and services at the national, European, and international level. Secondly, there is no coherent, general, or universal international regulation of AI or AI-driven products and services.

Nevertheless, even outside the EU there is widespread agreement that there is the need to have proportional and robust regulation in place, at least for high-risk AI-driven products and such services. If we look at the multiple fields where AI-driven systems are currently used and could be used in the future and also look closely at the inherent benefits and risks linked to those systems and products it seems less surprising that prominent heads of companies selling AI-driven products have emphasized the urgent need to regulate AI systems, products, and services, as well.[82]

The vulnerability of automated trading systems on the financial market may serve as an example highlighting the huge impact of intelligent systems: In the Flash Crash 2010, a quickly completed order triggered automated selling, wiping out nearly $1,000 billion worth of US shares for a period of several minutes.[83]

Therefore, we agree with those who argue that high-risk AI products and such services are emerging and disruptive technologies that have to be regulated.[84] This is especially true with regard to high-risk AI services because these are often ignored. In our view, there is an urgent need for responsible, (i.e. robust) and proportional regulation of high-risk AI products and services today, because if we try to regulate these when major damages have already occurred, it will be too late.

[81] Critical on this as well C Wendehorst, Chapter 12, in this volume.

[82] This is true, for example, *Bill Gates*, *Sundar Pichai*, and *Elon Musk* have called for the regulation of AI. See S Pichai, 'Why Google Thinks We Need to Regulate AI' *Financial Times* (20 January 2020) www.ft.com/content/3467659a-386d-11ea-ac3c-f68c10993b04; E Mack, 'Bill Gates Says You Should Worry About Artificial Intelligence' (*Forbes*, 28 January 2015) www.forbes.com/sites/ericmack/2015/01/28/bill-gates-also-worries-artificial-intelligence-is-a-threat/; S Gibbs, 'Elon Musk: Regulate AI to Combat 'Existential Threat' before It's Too Late' *The Guardian* (17 July 2017) www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo: *Musk* stated in July 2017, at a meeting of the US National Governors Association, that 'AI is a fundamental risk to the existence of human civilization.'

[83] *Cf.* M Mackenzie and A van Duyn, '"Flash Crash" was Sparked by Single Order' *Financial Times* (1 October 2010) www.ft.com/content/8ee1a816-cd81-11df-9c82-00144feab49a. *Cf.* J Tallinn and T Ngo, Chapter 2, in this volume; M Paul, Chapter 21, in this volume.

[84] *Cf.* House of Lords Select Committee on Artificial Intelligence, *AI in the UK: Ready, Willing and Able?* (Report of Session 2017–2019, 2018) HL Paper 100, 126 *et seq.*; MU Scherer 'Regulating Artificial Intelligence: Risks, Challenges, Competencies, and Strategies' (2016) 29(2) *Harvard Journal of Law & Technology* 353, 355; Perri 6, 'Ethics, Regulation and the New Artificial Intelligence, Part I: Accountability and Power' (2010) 4 *INFO, COMM & SOC'Y* 199, 203.

V. A NEW APPROACH: ADAPTIVE REGULATION OF AI-DRIVEN HIGH-RISK PRODUCTS AND SERVICES

1. *A New Approach*

We argue that a new approach to regulating AI-driven products is important to avoid the shortfalls of the rules at the national, supranational, and international level mentioned earlier. Our aim is to establish a regulatory approach that can supplement preventive procedures and, at the same time, close the gaps of liability-based approaches of different legal systems. This approach shall be applicable universally and could be laid down in national, supranational, or international laws. Our proposal aims for a proactive, adaptive regulatory scheme that is flexible, risk-sensitive, and has the incentive to assess and lower risks by those companies that develop and sell high-risk AI-driven products and such services. The proposal's core is that an operator or company must pay a proportionate amount of money (called regulatory capital in the following) as a financial security for future damages before a high-risk, AI-based product or such a service enters the market. To avoid over-regulation, we focus on AI-based products belonging to a class of high-risk products and services which, accordingly, have the potential to cause major damages for protected individual values, rights or interests, or common goods, such as life and bodily integrity, the environment, or the financial stability of a State. A regulatory framework for the potential development of superhuman AI will be discussed as well.

The special case of autonomous weapons, also a high-risk product, has to be mentioned as well: With regard to the specific problems of the development of (semi-)autonomous weapons, many authors and States state, based on convincing arguments, that a prohibition of these weapons is mandatory due to ethical and legal considerations.[85] This could mean that any kind of adaptive regulation suggested here should not be discussed as such regulation could be a safety net and justify the market entry of such weapons. We agree with the former, that a prohibition of such weapons is feasible, but disagree with the latter. Our argument for including (semi-)autonomous weapons in this discussion about responsible and adaptive regulation does not mean that we endorse the development, production, or selling of (semi-)autonomous weapons – quite to the contrary. Currently, however, it seems unlikely that there will be a consensus by the relevant States that develop, produce, or sell such weapons to sign an international treaty prohibiting or limiting these products in a meaningful way.[86] Therefore, this chapter's proposed regulatory approach

[85] As, for instance, the government of Austria, *cf.* Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 'Proposal for a Mandate to Negotiate a Legally-Binding Instrument that Addresses the Legal, Humanitarian and Ethical Concerns Posed by Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS)' (Working Paper Submitted to the Convention on Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems by Austria, Brazil, and Chile, 8 August 2018) CCW/GGE.2/2018/WP.7 https://undocs.org/CCW/GGE.2/2018/WP.7; and *cf.* the decision of the *Österreicherischen Nationalrat*, Decision to Ban Killer Robots, 24 February 2021, www.parlament.gv.at/PAKT/VHG/XXVII/E/E_00136/index.shtml#.

[86] For the different State positions, see Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 'Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW)' (Report of the 2019 session of the GGE on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 25 September 2019) CCW/GGW.1/2019/3 https://undocs.org/en/CCW/GGE.1/2019/3. On the discussion of these views *cf.* Voeneky, 'Key Elements of Responsible Artificial Intelligence' (n 9) 15–16. *Cf.* as well the resolution of the European Parliament, EP Resolution of 20 October 2020 with recommendations to the Commission on a framework for ethical aspects of artificial intelligence, robotics, and related technologies (2020/2012(INL)) www.europarl.europa.eu/doceo/document/TA-9-2020-10-20_DE.html#sdocta8.

could, and should, at least close the responsibility gap that emerges if such weapons are developed and used. This seems to be urgently necessary as there are lacunae in the traditional rules of international humanitarian law,[87] and international criminal law,[88] and the international rules on State responsibility.[89] There is the danger that, because of these lacunae, States do not even have to pay compensation if, for instance, an autonomous weapon is attacking and killing civilians in clear violation of the rules of international law.

2. *Key Elements of Adaptive Regulation of AI High-Risk Products and Services*

We argue that adaptive regulation as a new regulatory scheme for AI-driven high-risk products and such services shall consist of the following core elements:

First, the riskiness of a specific AI-driven product or service should be evaluated by a commission of independent experts. The threshold regarding whether such an evaluation has to take place is dependent on whether the AI-based product or service falls into a high-risk category according to a *prima facie* classification of its riskiness that shall be laid down in legal rules.[90] Possible future scenarios together with available data on past experiences (using the evaluated or similar products or services) will form the basis for the experts' evaluation. If the evaluated product or service is newly developed, a certain number of test cases proposed by the expert commission should provide the data for evaluation.

Second, after the expert commission has evaluated whether a specific AI-driven product or service is high-risk as defined above and falls under the new regulatory scheme, and the questions are answered in the positive, the expert committee shall develop risk scenarios that specify possible losses and associated likelihoods for the scenarios to realize.

Third, relying, in addition to the riskiness of the product, on the financial situation of the developing or producing company,[91] the experts will determine the specific regulatory capital that has to be paid. They shall also spell out an evaluation system that will allow measurement and assessment of future cases for damages due to the implementation or operation of the AI-driven product or service.

Fourth, the set-up of a fund is necessary, into which the regulatory capital has to be paid. This capital shall be used to cover damages that are caused by the AI-driven high-risk product or service upon occurrence. After a reasonable time, for instance 5–10 years, the capital shall be paid back to the company if the product or service has caused no losses or damages.

Fifth, as mentioned above, after a high-risk product or service has entered the market, the company selling the product or service has to monitor the performance and effects of the product or service by collecting data. This should be understood as a compulsory monitoring phase in which monitoring schemes are implemented. The data will serve as an important source for future evaluation of the riskiness of the product by the expert commission.

[87] See Geneva Conventions (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31, 85, 135, 287; Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3; Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II) (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 609.

[88] Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3.

[89] ILC, 'Materials on the Responsibility of States for Internationally Wrongful Acts' (United Nations, 2012) ST/LEG/SER.B/25.

[90] For a proposal by the EU Commission, cf. Section II.

[91] In contrast, the Draft EU AIA obliges 'providers' and 'users', see Section IV 3 b).

In particular, if the product or service is new and data is scarce, the evaluation system is of utmost importance because it serves as a database for future decisions on the amount of the regulatory capital and on the need for future monitoring of the product or service.

Sixth, another element of the proposed governance scheme is that the company should be asked to develop appropriate test mechanisms. A testing mechanism is a valid and transparent procedure ensuring the safety of the AI-driven product. For instance, a self-driving vehicle must pass a sufficient number of test cases to ensure that these vehicles behave in a safe way, meeting a reasonable benchmark.[92] Such a benchmark and test mechanism should be determined by the expert commission. Market entry should not be possible without a test mechanism in place. Given the data from the monitoring phase, the expert commission will be able to evaluate the product; but an appropriate test mechanism has additional advantages as the company itself can use it for the continuous evaluation of the product. It can support the re-evaluation explained in the next step. It will also help the regulator provide automized test mechanisms for the monitoring and evaluating of the technology, particularly in similar scenarios.

Seventh, the expert commission shall re-evaluate the AI-driven high-risk product or service on a regular basis, possibly every year. It can modify its decision on the proportionate amount of regulatory capital that is needed to match the risks by relying on new information and assessing the collected data. The established evaluation system mentioned above will provide reliable data for relevant decisions. (And, as mentioned earlier, after a reasonable time frame, the capital should be paid back to the company if the product or service has caused no losses or damages.)

3. *Advantages of Adaptive Regulation*

The following significant advantages follow from the adaptive approach[93] to regulation of AI high-risk prodocts and services: It avoids over-regulating the use of AI products and services especially in cases if the AI technology is new, and the associated risks are *ex ante* unclear. Current regulatory approaches that lay down preventive permit procedures can prevent a products' market entry (if the threshold is too high) or allow the market entry of an unsafe product (if the threshold is too low or is not implemented). With the adaptive regulation approach, however, it will be possible to ensure that a new AI product or AI-based service enters the market while sufficient regulatory capital covers possible future damages. The capital will be paid back to the company if the product or service proves to be a low-risk product or service after an evaluation period by using the data collected during this time according to the evaluation system.

a. Flexibility

The adaptive regulation approach allows reacting fast and in a flexible way to new technological developments in the field of AI. Since only the regulation's core elements are legally fixed *a priori*, and details shall be adapted on a case-by-case basis by an expert commission, the specific framing for an AI (*prima facie*) high-risk product can be changed depending on the information and data available. A periodical re-evaluation of the product or service ensures that new information can be taken into account, and the decision is based on the latest data.

[92] See, for example, T Menzel, G Bagschik, and M Maurer, 'Scenarios for Development, Test and Validation of Automated Vehicles' (2018) IEEE Intelligent Vehicles Symposium (IV).

[93] For the notion of adaptive governance *cf.* Tate and Banda, 'Proportionate and Adaptive Governance' (n 27) 4 *et seq.*, 20.

b. Risk Sensitiveness

The approach is not only risk-sensitive with regard to the newly developed high-risk AI-based product or service; it also takes into account the different levels of risks accepted by different societies and legal cultures. It can be assumed that different States and societies are willing to accept different levels of risks linked to specific AI products and services, depending on the expected benefit. If, for instance, a society is particularly dependent on autonomous vehicles because of an ageing population and deficits in the public transport system, it might decide to accept higher risks linked to these vehicles to have the chance of an earlier market entry of the AI-based cars. According to these common aims, the threshold to enter the market laid down as part of a permit procedure could be lowered if, at the same time, the regulatory capital will be paid in the fonds and ensures that (at least) all damages will be compensated. The same is true, for instance, for AI-driven medical devices or other AI high-risk products that might be particularly important to people from one State and the common good of specific society due to certain circumstances.

c. Potential Universality and Possible Regionalization

Nevertheless, as AI systems are systems that could be used in every part of the world, the expert commission and its decision shall be based on international law. An international treaty, incorporating the adaptive regulation approach into international law, could outbalance lacunae or hurdles based on national admission procedures that might be ineffective or insufficient. The commission's recommendations or decisions, once made public, could be implemented directly in different national legal orders if the risk sensitiveness of the State is the same, and could serve as a supplement for the national admission process.

If, however, different types of risk attitudes towards an AI-driven high-risk product or such a service in different States exist, a cultural bias of risk averseness (or risk proneness) can be taken into account when implementing the proposal for regulation spelled out in this chapter at the national or regional levels. This allows the necessary flexibility of a State to avoid insufficient regulation (or overregulation) whilst protecting individual rights, such as bodily integrity or health, or promoting the common good, as the environment or the financial stability of a State or region. Such adjustments can be deemed necessary, especially in democratic societies, if risk perception of the population changes over time, and lawmakers and governments have to react to the changed attitudes. To that end, the German Constitutional Court (*Bundesverfassungsgericht*, BVerfG) has held that high-risk technologies (in the case at hand: nuclear energy) are particularly dependent on the acceptance of the population in the democratic society, because of the potentially severe damages that might be caused if they are used. The Constitutional Court stressed that because of a change in the public's perception of a high-risk technology, a reassessment of this technology by the national legislator was justified – even if no new facts were given.[94]

d. Monitoring of Risks

It can be expected that in most cases, a company producing a high-risk AI-driven product or service will be *a priori* convinced of the safety of its product or service and will argue that its AI-driven product or service can be used without relevant risks, while this opinion is possibly not

[94] BVerfGE 143, 246–396 (BVerfG *1 BvR 2821/11*) para 308. One of the questions in the proceedings was whether the lawmaker in Germany can justify the nuclear phase-out that was enacted after the reactor accident in Fukushima, Japan, took place. This was disputed as an 'irrational' change of German laws as the reactor accident in Fukushima did not, in itself, change the risk factors linked to nuclear reactors located in Germany.

shared by all experts in the field. Therefore, the collection of data on the product's performance in real-world settings by the company evaluation systems is an important part of the adaptive regulation proposal introduced in this chapter. On the one hand, the data can help the company to show that its product or service is, as claimed, a low-risk product after a certain evaluation period and justify that the regulatory capital could be reduced or paid back; on the other hand, if the AI-driven product causes damages, the collected data will help improve the product and remedy future problems of using the technology. The data can also serve as an important source of information when similar products have to be evaluated and their risks have to be estimated. Hence, a monitoring phase is an important element of the proposal as reliable data are created on the product's or service's performance, which can be important at a later stage to prove that the technology is actually as riskless as claimed by the company at the beginning.

e. Democratic Legitimacy and Expert Commissions

The adaptive regulation approach spelled out in this chapter is not dependent on the constitution of a democratic, human rights-based State, but it is compatible with democracy and aims to protect core human and constitutional rights, such as life and health, as well as common goods, such as the environment. In order to have a sufficient basis that is legitimized, the rules implemented by the expert commission and the rules establishing the expert commission, should be based on an Act of parliament. Legally enshrined expert commissions or panels already exist in different contexts as part of the regulation of disruptive, high-risk products or technologies. They are a decisive element of permit procedures during the development of new drugs, as laid down for instance in the German Medicinal Products Act (*Arzneimittelgesetz*).[95] Another example of an interdisciplinary commission based on an act of parliament is the area of biotechnology regulation in Germany.[96]

As long as the commission's key requirements, such as the procedure for the appointment of its members, the number of members, the scientific background of members, and the procedure for the drafting of recommendations and decisions, are based on an act of parliament, a sufficient degree of democratic legitimacy is given.[97] In a democracy, this will avoid the pitfalls of elitism and an expert system, an expertocracy, that does not possess sufficient links to the legislature of a democratic State. A legal basis further complies with the requirements of human and constitutional rights-based constitutions, such as the German Basic Law, which demand that the main decisions relevant for constitutional rights have to be based on rules adopted by the legislative.[98]

[95] §§ 40(1), 42(1) AMG (n 53). For details *cf.* S Voeneky, *Recht, Moral und Ethik* (2010) 584–635, esp. at 594–606 (hereafter S Voeneky, *Recht, Moral und Ethik*).

[96] See the Central Committee on Biological Safety (ZKBS), an expert commission responsible for evaluating the risks concerning the development and use of genetically modified organisms (GMOs) www.zkbs-online.de/ZKBS/EN/Home/home_node.html. The commission is based on the the German Genetic Engineering Act (*Gentechnikgesetz* (GenTG)); BGBl 1993 I 2066 (§ 4 GenTG) and the decree, *Verordnung über die Zentrale Kommission für die Biologische Sicherheit* (ZKBS-Verordnung, ZKBSV) 30 October 1990 www.gesetze-im-internet.de/zkbsv/index.html.

[97] S Voeneky, *Recht, Moral und Ethik* (n 98).

[98] The so-called *Wesentlichkeitsprinzip*, that can be deduced from German Basic Law, is dependent on the constitutional framing and is not a necessary element of every liberal human rights-based democracy. In the United States, for instance, it is constitutional that the US president issues Executive Orders that are highly relevant for the exercise of constitutional rights of individuals, without the need to have a specific regulation based on an act of parliament. For the 'Wesentlichkeitsprinzip' according to the German Basic Law *cf.* S Voeneky, *Recht, Moral und Ethik* (2010) 214–218 with further references; B Grzeszick, 'Art. 20' in T Maunz und G Dürig (eds), *Grundgesetz-Kommentar* (August 2020) para 105.

f. No Insurance Market Dependency

The adaptive regulation approach spelled out in this chapter avoids reliance on a commercial insurance scheme. An approach that refers to an insurance scheme that obliges companies to procure insurance for their AI-based high-risk products or services would depend on the availability of such insurances from companies. This could, however, fail for practical or structural reasons. Further, insurance might not be feasible for the development of new high-risk AI products and services if, and because, only a limited amount of data is available.[99] Besides, low probability-high-risk scenarios with unclear probability can hardly be covered adequately by insurances, as risk-sharing might be impossible or difficult to achieve by the insurer. Lastly, the reliance on insurance would mean that higher costs have to be covered by a company that is producing AI-based products, as the insurance company needs to be compensated for their insurance product and aims to avoid financial drawbacks by understating risks.

At the national level, there is an example that an attempt to regulate a disruptive technology, in this case biotechnology, based on the duty to get insurance failed as this duty was not implemented by either the regulator or the insurance industry.[100] Even at the international level, the duty to get insurance for operators can be seen as a major roadblock for ratifying and implementing an international treaty on the liability for environmental damage.[101]

4. *Challenges of an Adaptive Regulation Approach for AI-Driven High-Risk Products*

a. No Financial Means?

A first argument against the adaptive regulation approach could be that (different from financial institutions) the companies that develop and sell disruptive high-risk AI products or services do not have the capital to pay a certain amount as a guarantee for possible future damages caused by the products or service. This argument is, on the one hand, not convincing if we think about well-established big technology companies, like Facebook, Google, or Apple, etc., that develop AI products and services or outsource these developments to their subsidiaries.

On the other hand, start-ups, and new companies might develop AI-driven products and services which fall within the high-risk area. However, these companies often receive funding capital from private investors to achieve their goals even if they generate profit at a very late stage.[102] If an investor, often a venture capitalist, knows that the regulatory requirement is to pay a certain amount of capital to a fund that serves as security but that capital will be paid back to the company after a

[99] This is the problem existing with regard to the duty to get insurance for an operator that risks causing environmental emergencies in Antarctica as laid down in the Liability Annex to the Antarctic Treaty (Annex VI to the Protocol on Environmental Protection to the Antarctic Treaty: Liability Arising from Environmental Emergencies (adopted on 14 June 2005, not yet entered into force), *cf.* IGP&I Clubs, *Annex VI to the Protocol on Environmental Protection to the Antarctic Treaty: Financial Security* (2019), https://documents.ats.aq/ATCM42/ip/ATCM42_ip101_e.doc.

[100] Pursuant to § 36 GenTG (n 96) the German Federal Government should implement the duty to get insurance with the approval of the Federal Council (*Bundesrat*) by means of a decree. Such a secondary legislation, however, has never been adopted, *cf.* Deutscher Ethikrat, *Biosicherheit – Freiheit und Verantwortung in der Wissenschaft: Stellungnahme* (2014) 264 www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-biosicherheit.pdf.

[101] *Cf.* the so-called Liability Annex, an international treaty, not yet in force, that regulates the compensation of damages linked to environmental emergencies caused by an operator in the area of the Antarctic Treaty System, see note 58.

[102] For example, *Tesla* as a car manufacturer trying to develop (semi-)autonomous cars has only generated profit since 2020, *cf.* 'Tesla Has First Profitable Year but Competition Is Growing' (*The New York Times*, 27 January 2021) www.nytimes.com/2021/01/27/business/tesla-earnings.html.

certain time if the product or service does not cause damages, this obligation would not impede or disincentivize the financing of the company compared to other requirements (for instance, as part of permit procedures). Quite to the contrary: To lay down a threshold of a certain amount of regulatory capital as a necessary condition before market-entry of an AI-based high-risk product (not for the stage of the research or development of the product) or AI-based service is an opportunity for the investor to take those risks into account that the company itself might downplay.

In the event that a State is convinced that a certain AI-driven product or service is fostering the common good of its society, and private investors are reluctant to finance the producing company because of major or unclear risks linked to the product or service, there is the possibility that the particular State may support the company with its financial means. Financial support has been given in different forms in other cases of the development of high-risk technology or products in the past and present.[103]

b. Ambiguity and Overregulation?

Another argument one could envisage against the adaptive regulatory approach introduced in this chapter is that it is unclear which AI-driven products or services have to be seen as high-risk products or high-risk services; and therefore there might be an inherent bias that leads to over-regulation as the category of high-risk products or services cannot be determined without grey areas, and can be determined neither precisely nor narrowly enough. However, what could be brought forward against this argument is that the category of high-risk AI products and services that the expert commission shall evaluate will be laid down in national, supranational, or international law after a process that includes the discourse with different relevant actors and stakholders, such as companies, developers, researchers, etc.[104] Criteria for a classification of prima facie high-risk AI products or services should be the possible damage that can occur if a certain risk linked to the product or service materializes. In order to avoid overregulation, one should limit the group of AI-driven high-risk products and services to the most evident; this might be depending on the risk proneness or risk awareness of a society as long as there is no international consensus.

c. Too Early to Regulate?

To regulate emerging technologies such as AI-based products and services is a challenge, and the argument is often brought forward that it is too early to regulate the technologies because the final product or service is unclear at a developmental stage. This is often linked to the argument that regulation of emerging technologies will mean inevitable overregulation of these technologies, as mentioned earlier. The answer to these arguments is that we as a society, every State, and the global community as a whole should avoid falling into the 'it is too early to regulate until it is too late' trap. Dynamic developments in a high-risk emerging technology sector, in particular, are characterized by the fact that sensible regulation rather might come too late, as legislative processes are, or can often be, lengthy. The advantage of the adaptive regulation proposed in this chapter is that, despite regulation, flexible standardization adapted to the specific case and the development of risk is possible.

[103] For instance, during the COVID-19 pandemic certain vaccine developing companies in Germany have been supported by the federal government and the EU; for example, the *Kreditanstalt für Wiederaufbau* (KfW) has acquired 'minority interest in CureVac AG on behalf of the Federal Government', *cf. KfW*, 6 August 2020. Also, the high-risk technology of nuclear power plants have been supported financially by different means in Germany since their establishment; *inter alia* the liability of the operating company in case of a maximum credible accident that has been part of the German law is capped and the German State is liable for compensation for the damages exceeding this cap, *cf.* §§ 25 *et seq.*, 31, 34, and 38 German Atomic Energy Act (*Atomgesetz* (AtG)), BGBl 1985 I 1565.

[104] *Cf.* above the proposals of the EU Parliament, note 35.

d. No Independent Experts?

As mentioned earlier, the inclusion of expert commissions and other interdisciplinary bodies, such as independent ethics committees and Institutional Review Boards, has been established in various areas as an important element in the context of the regulation and assessment of disruptive, high-risk products or procedures. There are no reasons to assume why expert commissions should not be a decisive and important element in the case of AI regulation. Transparency obligations might ensure that experts closely linked to certain companies are not part of such a commission or are not part of a specific decision of such a commission. Moreover, a pluralistic and interdisciplinary composition of such a body is able to prevent biases as part of the regulative process.[105]

e. Unacceptable Joint Liability of Companies?

Further, it is not an argument against the fund scheme that companies that distribute AI-based products or services that later turn out to be low-risk are unduly held co-liable for companies that produce and distribute AI-based products or services that later turn out to be high-risk and cause damage. The aim of the fund's establishment is that claims for damages against a certain company X are initially compensated from the fund after a harmful case, namely from the sum that the harm-causing company X has deposited precisely for these cases concerning its risky AI products and services; should the amount of damage exceed this, further damages should initially be paid by company X itself. Thus, unlike with funds that contain a total capital that is depleted when damage payments are made in large amounts, it would be ensured that, in principle, the fund would continue to exist with the separate financial reserves of each company. If, to the contrary, the entire fund would be liable in the event of damage, the state where the company Y producing low-risk AI products is a national would have to provide a default liability to guarantee the repayment of the capital to the company Y. The state would be obliged to reimburse the paid-in regulatory capital to a company such as Y if, contrary to expert opinion, an AI product turns out to be low-risk and the regulatory capital has to be repaid to the company, but the fund does not have the financial means to do so due to other claims.

VI. DETERMINING THE REGULATORY CAPITAL

Central to the adaptive regulation proposed here is determining the level of regulatory capital. In this Section, we provide a formal setup, using probabilistic approaches.[106] In the first example, we consider a company that may invest in two competing emerging AI-driven products; one of the products is substantially riskier than the other. Even if we presume that the company is acting rationally (in the sense of a utility maximising[107] company),[108] there are good reasons to claim that risks exceeding the assets of the company will not be taken fully into account in the decision process of this company because, if the risks materialize, the bankruptcy of the company will be caused. Although it seems *prima facie* rational that diminishing risks exceeding the assets of the

[105] In the area of biotechnology *cf.* for instance in Germany the Central Committee on Biological Safety, ZKBS, note 96.

[106] *Cf.* VV Acharya and others, 'Measuring Systemic Risk' (2017) 30(1) *The Review of Financial Studies* 2–47 (hereafter Acharya and others, 'Measuring Systemic Risk').

[107] For this initial claim it is not necessary that utility is measured on a monetary scale. Later, when it comes to determining regulatory capital, we will, however, rely on measuring utility in terms of wealth.

[108] This means that future profits and losses are weighted with a utility function and then averaged by expectation. See for example DM Kreps, *A Course in Microeconomic Theory* (1990) or A Mas-Colell, MD Whinston, and JR Green, *Microeconomic Theory* (1995) volume 1.

company should be the priority for the management of a company, as these risks threaten this actor's existence, the opposite behavior is incentivized. The high or even existential risks will be neglected by the company if there is no regulation in place obliging the company to take them into account: The company will seek high-risk investments because the higher return is not sufficiently downweighed by expected losses, which are capped at the level of the initial endowment.[109]

First Example: Two competing AI technologies or products

Consider a company with an initial endowment w_0. The company can decide to invest in two different AI-driven products or technologies offering (random) returns r and r' for the investment of 1 unit of currency. The first technology is the less risky one, while the second is riskier. We assume there are two scenarios: The first scenario (the best case, denoted by $+$) is if the risk does *de facto* not materialize. This scenario is associated with some probability p. In this scenario, the riskier strategy offers a higher return, i.e. $r(+) < r'(+)$.

In the second scenario (the worst case, denoted by $-$ and having probability $1-p$), the riskier technology will lead to larger losses, such that we assume $0 > r(-) > r'(-)$, both values being negative (yielding losses).

Summarizing, when the company invests the initial endowment into the strategy, the wealth at the end of the considered period (say at time 1) will be $w_1 = w_0 \cdot r$, on investing in the first technology, or $w_1' = w_0 \cdot r'$, when investing in the second, riskier technology, bankruptcy will occur when $w_1 < 0$, or $w_1' < 0$, respectively.

We assume that the company maximizes expected utility: Expected utility of the first strategy is given by the expectation of the utility of the wealth at time 1, $EU = E\left[u(w_1)1_{\{w_1>0\}}\right]$ (or $EU' = E\left[u(w_1')1_{\{w_1'>0\}}\right]$, respectively for the second strategy). Here u is a utility function[110] (we assume it is increasing), E denotes the expectation operator, and $1_{\{w_1>0\}}$ is the indicator function, being equal to one if $w_1 > 0$, (no bankruptcy) and zero otherwise (and similarly $1_{\{w_1'>0\}}$). The company chooses the strategy with the highest expected utility, namely, the first one if $EU > EU'$ and the second one if $EU' > EU$. If both are equal, one looks for additional criteria to find the optimal choice. This is typically a rational strategy.

Up to now, we have considered a standard case with two scenarios, a best case and a worst case. In the case of emerging and disruptive technologies, failure of high-risk AI systems and AI-driven products might lead to immense losses, such that in the worst-case scenario ($-$) bankruptcy occurs. This changes the picture dramatically:

> we obtain that $EU = p \cdot u(w_0 \cdot r(+))$ for the first technology, and for the second, riskier technology $EU' = p \cdot u(w_0 \cdot r'(+))$. Since the riskier technology's return in the best case scenario is higher, the company will prefer this technology. Most importantly, this does neither depend on the worst case's probability nor on the amount of the occurring losses. The company, by maximizing utility, will not consider losses beyond bankruptcy in its strategy.

Summarizing, the outcome of this analysis highlights the importance of regulation in providing incentives for the company to avoid overly risky strategies.

[109] See E Eberlein and DB Madan, 'Unbounded Liabilities, Capital Reserve Requirements and the Taxpayer Put Option' (2012) 12(5) *Quantitative Finance* 709–724 and references therein.

[110] A utility function associates to a various alternative a number (the utility). The higher the number (utility) is, the stronger the alternative is preferred. For example, 1 EUR has a different value to an individual who is a millionaire in comparison to a person who is poor. The utility function is able to capture such (and other) effects. See H Föllmer and A Schied, *Stochastic Finance: an Introduction in Discrete Time* (2011) Chapter 2 for further references.

The first example highlights that a utility-maximising company will accept large risks surprisingly easily. In particular, the exact amount of losses does not influence the rational decision process, because losses are capped at the level of bankruptcy and the hypothetical losses are high enough to lead to bankruptcy regardless. It can be presumed that the company does not care about the particular amount of losses once bankruptcy occurs. This, in particular, encourages a high-risk strategy of companies since strategies with higher risk on average typically promise higher profits on average. However, the proposed adaptive regulation can promote the common good in aiming to avoid large losses. We will show below that the proposed regulation brings large losses back into the utility maximization procedure by penalizing high losses with high regulative costs, thus helping to avoid these.

Considering the problem of superhuman AI, a particular challenge arises: Once a company develops superhuman AI, the realized utility will be huge. It is argued that a superhuman AI cannot be controlled; thus, it is posing an existential threat not restricted to the company. Potential losses are clearly beyond any scale, yet any company will aim to develop such a superintelligent system as the benefits will be similarly beyond any scale.

The example highlights that a need for regulation will hopefully provide guidance for controlling the development of such AI systems when high-risk AI products lead to large losses and damages. However, with a low or even very low probability of this, large losses, once occurred, have to be compensated for by the public, since the company will be bankrupt and no longer able to cover them. Hence, regulation is needed to prevent a liability shortfall.

The following example will show that a reasonable regulation fosters an efficient maximization of overall wealth in comparison to a setting without regulation.

Second Example: A stylized framework for regulation

In this second example, regulatory capital is introduced. Adaptive regulation can maximize the overall wealth, minimize relevant risks, avoid large losses and foster the common good by requiring suitable capital charges.

Consider I companies: each company i has an initial wealth $\bar{w}_0^i$, where one part $\bar{w}_0^i - w_0^i$ is consumed initially, and the other part w_0^i is invested (as in the above example) resulting in the random wealth w_1^i at time 1. The company i pays a regulatory capital ρ^i and, therefore, aims at the following maximization:

$$max\Big[c \cdot \left(\bar{w}_0^i - w_0^i - \rho^i\right) + E\Big[u\Big(w_1^i 1_{\{w_1^i > 0\}}\Big)\Big]\Big]$$

The relevant rules should aim to maximize overall wealth: In the case of bankruptcy of a company, say i, the public and other actors have to cover losses. We assume that this is proportional to the occurred losses, $g \cdot w_1^i 1_{\{w_1^i < 0\}}$. The overall welfare function $P^1 + P^2$ consists of two parts: the first part is simply the sum of the utility of the companies,

$$P^1 = \sum_{i=1}^{I} c \cdot \left(\bar{w}_0^i - w_0^i - \rho^i\right) + E\Big[u\Big(w_1^i 1_{\{w_1^i > 0\}}\Big)\Big].$$

The second part,

$$P^2 = \sum_{i=1}^{I} E\Big[g \cdot w_1^i 1_{\{w_1^i < 0\}}\Big],$$

is the expected costs in case of bankruptcies of the companies. As scholars argue,[111] one obtains the efficient outcome, maximizing overall wealth or the common good, respectively, by choosing regulatory capital as

$$\rho^i = \frac{g}{c} \cdot P\left(w_1^i < 0\right) \cdot ES^i; \tag{1}$$

here the expected shortfall is given by $ES^i = -E\left[w_1^i 1_{\{w_1^i<0\}}\right]$. Hence, by imposing this regulatory capital, the companies will take losses beyond bankruptcy into account, which will help to achieve maximal overall wealth.

As spelled out in the literature, one could incorporate systemic effects in addition, which we do not consider here for simplicity.[112]

Here the adaptive regulatory approach relies on expectations and, therefore, assumes that probabilities can be assessed, even if they have to be estimated[113] or suggested by a team of experts. In the case of high uncertainty, this might no longer be possible, and one can rely on non-linear expectations (i.e. utilize *Frank Knight's* concept of uncertainty or in the related context of 'uncertain futures'). As already mentioned, the projection of unknown future risks can be formalized by relying on extreme value theory.[114] Therefore, it is central that adapted methods are used to incorporate incoming information resulting from the above mentioned monitoring process or other sources. The relevant mathematical tools for this exist.[115]

VII. DISSENT AND EXPERT COMMISSION

With regard to the expert commission, one has to expect that a variety of opinions arise. One possibility is that the worst-case opinion is considered, that is, taking the most risk-averse view. An excellent alternative to taking best-/worst-case scenarios or similar estimates is to rely on the underlying estimates' credibility. This approach is based on the so-called credibility theory, which combines estimates, internal estimates, and several expert opinions in the actuarial context.[116] We show how and why this is relevant for the proposed regulation.

[111] Acharya and others, 'Measuring Systemic Risk' (n 109).
[112] Acharya and others, ibid.
[113] M Pitera and T Schmidt, 'Unbiased Estimation of Risk' (2018) 91 *Journal of Banking & Finance* 133–145.
[114] See, for example L De Haan and A Ferreira, *Extreme Value Theory: An Introduction* (2007).
[115] See, for example AH Jazwinski, *Stochastic Processes and Filtering Theory* (1970), R Frey and T Schmidt, 'Filtering and Incomplete Information' in T Bielecki and D Brigo (eds), *Credit Risk Frontiers* (2011); T Fadina, A Neufeld, and T Schmidt, 'Affine Processes under Parameter Uncertainty' (2019), 4.1 *Probability, Uncertainty and Quantitative Risk*, 1–35.
[116] Credibility theory refers to a *Bayesian* approach to weight the history of expert opinions, see the recent survey by R Norberg (2015) 'Credibility Theory' in N Balakrishnan and others (eds) *Wiley StatsRef: Statistics Reference Online* or the highly influential work by H Bühlmann, 'Experience Rating and Credibility Theory' (1967) 4(3) *ASTIN Bulletin* 199.

Third Example: Regulation relying on credibility theory

For simplicity, i will be fixed, and we consider only two experts, one suggesting the probability P_1 and the other one P_2. The associated values of the regulatory capital computed using equation (1) are denoted by ρ_1 and ρ_2, respectively.

The idea is to mix ρ_1 and ρ_2 for the estimation of the regulatory capital as follows:

$$\rho^{credible}(\theta) = \theta \cdot \rho_1 + (1-\theta) \cdot \rho_2$$

where θ will be chosen optimal in an appropriate sense. If we suppose that there is already experience on estimates of the two experts, we can obtain variances v_1 and v_2 estimated from their estimation history. The estimator having minimal variance is obtained by choosing

$$\theta_{opt} = \frac{v_2}{v_1 + v_2}.$$

When expert opinions differ, credibility theory can be used to provide a valid procedure for combining the proposed models. Systematic preference is given to experts who have provided better estimates in the past. Another alternative is to select the estimate with the highest (or lowest) capital; however, this would be easier to manipulate. More robust variants of this method based on quartiles, for example, also exist.

VIII. SUMMARY

This chapter spells out an adaptive regulatory model for high-risk AI products and services that requires regulatory capital to be deposited into a fund based on expert opinion. The model allows compensating potentially occurring damage, while at the same time motivating companies to avoid major risks. Therefore, it contributes to the protection of individual rights of persons, such as life and health, and to the promotion of the common good, such as the protection of the environment. Because regulatory capital is reimbursed to a company if an AI high-risk product or service is safe and risks do not materialize for years, we argue that this type of AI regulation will not create unnecessarily high barriers to the development, market entry, and use of new and important high-risk AI-based products and services. Besides, the model of adaptive regulation proposed in this chapter can be part of the law at the national, European, and international level.

9

China's Normative Systems for Responsible AI

From Soft Law to Hard Law

Weixing Shen and Yun Liu

I. INTRODUCTION

Progress in Artificial Intelligence (AI) technology has brought us novel experiences in many fields and has profoundly changed industrial production, social governance, public services, business marketing, and consumer experience. Currently, a number of AI technology products or services have been successfully produced in the fields of industrial intelligence, smart cities, self-driving cars, smart courts, intelligent recommendations, facial recognition applications, smart investment consultants, and intelligent robots. At the same time, the risks of fairness, transparency, and stability of AI have also posed widespread concerns among regulators and the public. We might have to endure security risks when enjoying the benefits brought by AI development, or otherwise to bridge the gap between innovation and security for the sustainable development of AI.

The Notice of the State Council on Issuing the Development Plan on the New Generation of Artificial Intelligence declares that China is devoted to becoming one of the world's major AI innovation centers. It lists four dimensions of construction goals: AI theory and technology systems, industry competitiveness, scientific innovation and talent cultivation, and governance norms and policy framework.[1] Specifically, by 2020, initial steps to build AI ethical norms and policies and legislation in related fields has been completed; by 2025, initial steps to establish AI laws and regulations, ethical norms and policy framework, and to develop AI security assessment and governance capabilities shall be accomplished; and by 2030, more complete AI laws and regulations, ethical norms, and policy systems shall be accomplished. Under the guidance of the plan, all relevant departments in Chinese authorities are actively building a normative governance system with equal emphasis on soft and hard laws.

This chapter focuses on China's efforts in the area of responsible AI, mainly from the perspective of the evolution of the normative system, and it introduces some recent legislative actions. The chapter proceeds mainly in two parts. In the first part, we would present the process of development from soft law to hard law through a comprehensive view on the normative system of responsible AI in China. In the second part, we set out a legal framework for responsible AI with four dimensions: data, algorithms, platforms, and application scenarios, based on statutory requirements for responsible AI in China in terms of existing and developing

[1] State Council, *The Notice of the State Council on Issuing the Development Plan on the New Generation of Artificial Intelligence* (*The State Council of the People's Republic of China*, 8 July 2017) www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

laws and regulations. Finally, this chapter concludes by identifying the trend of building a regulatory system for responsible AI in China.

II. THE MULTIPLE EXPLORATION OF RESPONSIBLE AI

1. *The Impact of AI Applications Is Regarded As a Revolution*

Science and technology are a kind of productive force. The innovation and application of new technologies often improves production efficiency and stimulates transformative changes on politics, economics, society, and culture. In China, 'technological revolution' got its name due to the widespread application of these technologies. It is well known that there have been three technological revolutions in the modern era. China missed the three historic developmental opportunities due to foreign invasion and internal turmoil. During the first and second industrial revolutions, which were powered by steam and electricity respectively, China was in its last imperial period, the Qing Dynasty, and missed the opportunity to participate in the creation of inventions because it was experiencing the century of humiliation in its five-thousand-year history. The Third Industrial Revolution, which began in the 1950s, was marked by the invention and application of atomic energy, electronic computers, space technology, and bioengineering. However, China missed most of it because it lacked the political environment to participate in international communications. China has been lagging behind for a long time. Due to the implementation of the reform and opening-up policy in 1978, China started to catch up and to learn from the West in the aspects of science and technology, legal systems, and other fields.

In order to promote the development of science and technology, Article 12 of the *Constitution of the People's Republic of China (1978 revised)* stipulates that the state shall vigorously develop scientific undertakings, strengthen scientific research, carry out technological innovation and technological revolution, and adopt advanced technology in all sectors of the national economy as far as possible. In September 1988, when *Deng Xiaoping*, the second generation leader of PRC, met with President *Gustáv Husák* of Czechoslovakia, he said, "Science and technology are the primary productive force," which has become a generally accepted consensus among Chinese people.

China caught up with the new trend of the third AI flourishing period. At the beginning of the twenty-first century, China's science and technology policy began to plan the development of 'next generation information technology'.[2] Since 2011, Chinese official documents have made extensive references to the development of 'next generation information technology'. With the rapid development of global AI, China has the opportunity to stand at the same starting line in the next round of AI technology development and application. China is fully aware of the profound impact of AI technology, and some high-level documents already refer to the next round of technological development, represented by AI, as a 'technological revolution', which is considered to be similar to the aforementioned three technological revolutions. As it is a revolutionary technology, the Chinese government does not see it only as a technology, but also realizes that it will play a key role in social governance, economic structure, political environment, the international landscape, and other aspects.

[2] Ministry of Science and Technology, 'Notice of the Project Proposal of Application for the National Key Research and Development Plan' (2001).

On 31 October 2018, the Political Bureau of the Central Committee of the CPC held its ninth collective study on the current status and trends of AI development, and *Xi Jinping* particularly emphasized that

> Artificial Intelligence is a strategic technology leading this round of scientific and technological revolution and industrial change, with a strong 'head goose' effect of spillover drive. It is necessary to strengthen the development of Artificial Intelligence potential risk research and prevention, to safeguard the interests of the people and national security, to ensure that Artificial Intelligence is safe, reliable and controllable. It is necessary to integrate multidisciplinary forces, strengthen research on legal, ethical and social issues related to AI, and establish and improve laws and regulations, institutional systems and ethics to safeguard the healthy development of AI.[3]

When recognizing that AI can have such a broad shaping power, China's technology policy reflects on the idea of balancing development and governance, considering both the promotion of positive social benefits from AI and the prevention of risks from AI applications as components of achieving responsible AI. On one hand, China's main goal, since its reform and opening up, has been to devote itself to economic development and the improvement of people's living standards, and in recent years it has also put forward the reform goal of modernizing its governance system and capabilities.[4] Actively promoting AI technology development is conducive to improving the country's economy, increasing people's well-being, and improving the social governance system. On the other hand, AI replaces or performs some behaviors on behalf of people with technical tools, and there is a risk of abuse or loss of control when the technical conditions and social situation are not yet mature. The development measures of technology and risk governance measures are two dimensions with large differences, and the responsible AI mentioned subsequently in this chapter focuses on analyzing the normative system of responsible AI in China from the risk governance dimension.

II. THE SOCIAL CONSENSUS ESTABLISHED BY SOFT LAW

Soft law is a common tool in the field of technology governance. Technical standards, ethics and morality, initiative and guidelines, and other forms of soft law have diverse flexibility and inclusiveness, and they can fill in areas of social relationships that hard law fails to adjust in a timely manner, adapting to the dual goals of technological innovation development and security

[3] *Xi Jinping* in the ninth collective study of the Political Bureau Central Committee of the CPC stressed the importance of strengthening leadership to do a good job of planning a clear task of solid foundation to promote the healthy development of a new generation of AI in China; Xinhua News Agency, 'Xi Jinping Presided Over the Ninth Collective Study of the Political Bureau of the CPC Central Committee and Gave a Speech' (*The State Council, The People's Republic of China*, 31 October 2018) www.gov.cn/xinwen/2018-10/31/content_5336251.htm (hereafter Xi Jinping, 'Ninth Study CPC Central Committee').

[4] In November 2013, the Third Plenary Session of the 18th CPC Central Committee took "promoting the modernization of national governance system and governance capacity" as the overall goal of comprehensively deepening reform, China.org.cn, 'Communiqué of the Third Plenary Session of the 18th Central Committee of the Communist Party of China' (*China.org.cn*, 15 January 2014) www.china.org.cn/china/third_plenary_session/2014-01/15/content_31203056.htm. On 31 October 2019, the Fourth Plenary Session of the 19th Central Committee of the Communist Party of China adopted the "decision of the Central Committee of the Communist Party of China on several major issues on adhering to and improving the socialist system with Chinese characteristics and promoting the modernization of national governance system and governance capacity", which further put forward the requirements of national governance reform, Online Party School, 'Communiqué of the Fourth Plenary Session of the 19th Central Committee of the Communist Party of China' (*Liaoning Urban and Rural Construction Planning Design Institute Co. LTD*, 5 December 2019) http://lnupd.com/english/article/shows/377.

prevention. In China's AI governance framework, government opinions, technical standards, and industry self-regulatory initiatives are all governance tools. These soft laws have no mandatory effect, but are mainly adopted and enforced through self-adoption, being referenced in contracts, within industry autonomy, through public opinion supervision, and within market competition to form a common social consciousness and be implemented, and other tools such as technical standards will also indirectly obtain binding effect by means of legal references.

A government opinion is a kind of nonmandatory guidance document issued by the government. In November 2017, the Ministry of Science and Technology of PRC led the establishment of the Office of the Development and Advancement of the New Generation of Artificial Intelligence, which is a coordinating body jointly composed of 15 relevant departments responsible for promoting the organization and implementation of the new generation of AI development planning and major science and technology projects. In March 2019, the Office of the Development and Advancement of the New Generation of Artificial Intelligence established the Committee on Professional Governance, which was formed by the Ministry of Science and Technology of PRC by inviting scholars from the fields of public administration, computer science, ethics, etc. On 17 June 2019, the Committee on Professional Governance of the New Generation of Artificial Intelligence released in its own name the *Governance Principles of the New Generation of Artificial Intelligence – Developing Responsible AI*.[5] According to the above governance principles, in order to promote the healthy development of a new generation of AI; better coordinate the relationship between development and governance; ensure safe, reliable, and controllable AI; promote sustainable economic, social, and ecological development; and build a community of human destiny; all parties involved in the development of AI should follow eight principles: (1) harmony and friendliness, with the goal of promoting common human welfare; (2) fairness and justice, eliminating prejudice and discrimination; (3) inclusiveness and sharing, in line with environmentally friendly, promoting coordinated development, eliminating the digital divide, and encouraging open and orderly competition; (4) respect for privacy, setting behavioral boundaries in the collection, storage, processing, use, and other aspects of personal information; (5) security and controllability, enhancing transparency, explainability, reliability, and controllability; (6) shared responsibility, clarifying the responsibilities of developers, users, and recipients; (7) open cooperation, encouraging interdisciplinary, cross-disciplinary, cross-regional, and cross-border exchanges and cooperation; (8) agile governance, ensuring timely detection and resolution of risks that may arise.[6] These principles establish the basic ethical framework for responsible AI in China.

China's technical standards include national standards, industry standards, and local standards which were published by governments agencies, and also include consortia standards and enterprise standards which were published by nongovernment agencies. According to the *Standardization Law of the People's Republic of China (2017 Revision)*, technical standards are in principle implemented voluntarily, and mandatory standards can be set only under specific circumstances.[7] There are no mandatory standards for AI governance, and those that have

[5] *Cf.* at www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html.

[6] The Committee on Professional Governance of the New Generation of Artificial Intelligence, 'Governance Principles of the New Generation of Artificial Intelligence – Developing Responsible AI' (*Catch the Wind*, 17 June 2019) www.ucozon.com/news/59733737.html.

[7] Article 10 of Standardization Law of the People's Republic of China (adopted 1988, effective 1989) stipulated that mandatory national standards shall be developed to address technical requirements for ensuring people's health and the security of their lives and property, safeguarding national and eco-environmental security, and meeting the basic need of economic and social management.

entered the work process are voluntary standards. In August 2020, the Standardization Administration of China and relevant departments released the *Guide to the Construction of National New Generation AI Standard System*, which incorporates security and ethics into the work plan of national standards, and plans to develop security and privacy protection standards, ethical standards, and other related standards.[8] In November 2020, the national information security standardization technical committee issued the *Guideline for Cyber Security Standards: Practice-Guideline for Ethics of Artificial Intelligence (Draft)*, which clearly lists five major types of ethical and moral risks of AI: (1) out-of-control risk, which is beyond the scope predetermined, understood, and controllable by the developer, designer, and deployer; (2) social risk, which causes social values and other systematic risks due to abuse or misuse; (3) infringement risk, which causes damage to basic rights, person, privacy, and property; (4) discrimination risk, which generates subjective or objective risks to specific groups of people; (5) liability risk, where the boundaries of responsibilities of relevant parties are unclear.[9] Currently, *AI Risk Assessment Model* and *AI Privacy Protection Machine Learning Technical Requirements* and other relevant technical standards have been released in draft version and are expected to become technical guidelines for AI risk assessment and privacy protection in the form of voluntary standards.[10]

Industry self-regulatory initiatives are nonbinding norms issued by a number of social groups and research institutions in conjunction with stakeholders. The Beijing Zhiyuan Institute of Artificial Intelligence, jointly built by Beijing's research institutions in the field of AI, released the *Beijing Consensus on Artificial Intelligence* in May 2019, which addresses AI from three aspects: research and development, use, and governance, and proposes 15 principles that are beneficial to the construction of a human destiny community and social development, which each participant should follow. In July 2021, AI Forum, jointly with more than 20 universities and AI technology companies, released the *Initiative for Promoting Trustworthy AI Development*, putting forward four initiatives: (1) insisting on technology for good to ensure that trustworthy AI benefits humanity; (2) insisting on sharing rights and responsibilities to promote the value concept of trustworthy AI; (3) insisting on a healthy and orderly approach to promote trustworthy AI industry practices; and (4) insisting on pluralism and inclusion to gather international consensus on trustworthy AI. In addition, there are a series of related initiative documents in areas such as facial recognition security.

III. THE AMBITION TOWARD A COMPREHENSIVE LEGAL FRAMEWORK

China currently does not have a unified AI law, but it has been under discussion. In contrast to soft law, the national legislature can promulgate a 'hard law' with binding force, which can establish general and binding rules on the scope of application, management system, security measures, rights and remedies, and legal liabilities of AI technologies. After these rules are confirmed by the legislator, the relevant actors within the scope of the law must implement a unified governance model. Therefore, by enacting laws, legislators are selecting a definitive model of governance for society. To ensure that the right choice is made, legislators need to have a good grasp of the past and present of the technology, as well as a sound understanding of the

[8] Standardization Administration of China, Cyberspace Administration of China, and other relevant departments, 'Guide to the Construction of National New Generation AI Standard System' (2020) 24–25.

[9] National Information Security Standardization Technical Committee, 'Guideline for Cyber Security Standards: Practice-Guideline for Ethics of Artificial Intelligence (Draft)' (2020).

[10] China Institute of Electronic Technology Standardization, 'White Paper on Standardization of Artificial Intelligence (version 2021)' (July 2021).

future direction of the technology. At the same time, in the early stages of the development of emerging technologies, there is a wide variation in the technological level of different developers, and the overall technological development stage of society is rapidly iterating, while the process of making new laws and revising them takes a long time, which leads legislators to worry that the laws made may soon become obsolete laws that lag behind the development stage of society, and that if there were no such laws made, it may face a series of new problems brought about by the development of disruptive innovations that cannot be clearly addressed.

During the two sessions of the National People's Congress in recent years, there have been many proposals or motions on AI governance. There are several proposals on AI regulation between 2018 and 2021, including the *Bill on Formulating the Law on the Development of Artificial Intelligence (2018)*, the *Bill on Formulating the Law on the Administration of Artificial Intelligence Applications (2019)*, and the *Bill on Formulating the Law on Artificial Intelligence Governance (2021)*. Other delegates have proposed the *Bill on the Enactment of a Law on Self-Driving Cars (2019)*. In accordance with the procedures of the two sessions of the National People's Congress, the delegates' bills will be referred to the relevant authorities for processing and response, mainly by the Legislative Affairs Commission of the Standing Committee of the National People's Congress, the Ministry of Science and Technology, and the Cyberspace Administration of China. At present, most of the proposals are referred to the legislative bodies or relevant industry authorities for research and solution, and their main attitude is that the AI legislation shall be carried out as a research project, not yet upgraded to the specific legislative agenda. For example, the Standing Committee of the National People's Congress (NPC) proposed in its 2020 legislative work plan to

> pay attention to research on legal issues related to new technologies and fields such as Artificial Intelligence, block chain and gene editing. Continue to promote the normalization and mechanism of theoretical research work, play the role of scientific research institutions, think tanks and other 'external brain', strengthen the exchange and cooperation with relevant parties, and urgently form high-quality research results.[11]

The legislative work on AI is also a task to which President *Xi Jinping* attaches importance. The Political Bureau of the CPC Central Committee held its ninth collective study on the current status and trends of AI development on October 31, 2018. The General Secretary of the CPC Central Committee and President *Xi Jinping* clearly stated at this meeting that China will, 'strengthen research on legal, ethical, and social issues related to Artificial Intelligence, and establish sound laws and regulations, institutional systems, and ethics to safeguard the healthy development of Artificial Intelligence.'[12] Subsequently, in November 2018, the members of the Standing Committee of the National People's Congress (NPC) held a special meeting in Beijing to discuss the topic of regulating the development of AI, and after discussion, it was concluded that

> the relevant special committees, working bodies and relevant parties of the NPC should take early action and act as soon as possible to conduct in-depth investigation and research on the

[11] Chinese National People's Congress, 'The 2020 legislative work plan of the Standing Committee of the National People's Congress (NPC)' (*The National People's Congress of the People's Republic of China*, 20 June 2020) www.npc.gov.cn/npc/c30834/202006/b46fd4cbdbbb4b8faa9487da9e76e5f6.shtml.

[12] Xi Jinping, 'Ninth Study CPC Central Committee' (n 3).

> legal issues involved in Artificial Intelligence, so as to provide relevant legislation work to lay a good foundation and make preparations to promote the healthy, standardized and orderly development of AI.[13]

During the two national sessions held in March 2019, more representatives and members began to discuss the topic of how to build the future rule of law system for AI.[14] In addition, according to the timetable established in the State Council's Development Plan for a New Generation of Artificial Intelligence, China should initially establish an AI legal and regulatory system in 2025. To this end, China's legislature has also begun to cooperate with experts from research institutions to conduct supporting studies. In this context, the author of this paper also participated in the relevant discussions, undertook one of the research tasks, and made suggestions on the legislative strategy of AI at the 45th biweekly consultation symposium of the 13th National Committee of the Chinese People's Political Consultative Conference (CPPCC) held in December 2020, undertook a project of the Ministry of Science and Technology in 2021 – Research on Major Legislative Issues of AI, and participated in the research task of the Law Working Committee of the Standing Committee of the National People's Congress on the legislation of facial recognition regulation.

Although there is no comprehensive legislative outcome, China's solutions for responsible AI can be extracted in all relevant laws. For example, the *E-Commerce Law of the People's Republic of China (E-Commerce Law)* enacted in 2018 dictates prohibitions on the use of personal information for Big-Data Driven Price Discrimination,[15] while the *Personal Information Protection Law of the People's Republic of China (Personal Information Protection Law)* and the *Data Security Law of the People's Republic of China (Data Security Law)* enacted in 2021 set requirements in terms of automated decision-making rules and data security requirements. In July 2021, the Supreme People's Court promulgated the *Provisions on Several Issues Concerning the Application of Law in Hearing Civil Cases Related to the Use of Facial Recognition Technology for Handling Personal Information* which is also an important governance regulation.[16] In addition, on 27 August 2021, the Cyberspace Administration of China issued the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)*, which is the first national-level legislative document in China to comprehensively regulate AI from the perspective of algorithms. At the local level, the Shenzhen legislature used its special legislative power as a special economic zone to issue the *Regulations on the Promotion of Artificial Intelligence Industry in the Shenzhen Special Economic Zone (Draft for Soliciting Public Comment)* on 14 July 2021. Despite the name of

[13] Li Zhanshu held and delivered speech in the meeting of the members of the Standing Committee of the National People's Congress: Xinhua, 'The Members of the NPC Standing Committee Chairman's Meeting Conducted Special Studies and Li Zhansu Chaired and Delivered a Speech' *(The National People's Congress of the People's Republic of China*, 24 November 2018) www.npc.gov.cn/npc/c238/201811/e3883fb5618e4a2bbefa5d170fe7b02a.shtml.

[14] Zhan Haifeng, *Committee Members Discuss about the Development of Artificial Intelligence: Building the Future Legal System of AI* (6th ed. 2019).

[15] Article 18 E-Commerce Law (promulgated 31 August 2018, effective 1 January 2019): when providing the results of search for commodities or services for a consumer based on the hobby, consumption habit, or any other traits thereof, the e-commerce business shall provide the consumer with options not targeting his/her identifiable traits and respect and equally protect the lawful rights and interests of consumers.

[16] Supreme People's Court, Law Interpretation [2021] No. 15, *Provisions on Several Issues Concerning the Application of Law in Hearing Civil Cases Related to the Use of Facial Recognition Technology for Handling Personal Information* (Judgement of 8 June 2021, in force on 1 August 2021) (hereafter Supreme People's Court, Provisions on Facial Recognition).

the law containing the word 'promotion', it includes a special chapter 'Governance Principles and Measures' providing the rules for responsible AI.

IV. THE LEGALLY BINDING METHOD TO ACHIEVE RESPONSIBLE AI

The new generation of AI is mainly driven by data and algorithms exerting essential social influence on different scenarios through various network platforms. Under the Chinese legal system, we can implement responsible AI through the governance of four dimensions: data, algorithm, platform, and application scenario.[17]

1. *Responsible AI Based on Data Governance*

Data is the key factor driving the prosperous development of a new generation of AI. Big data resources are increasingly having a significant impact on global production, circulation, distribution, consumption, and economic and social systems, as well as national governance capabilities.[18] The *Cyber Security Law* enacted in November 2016 sets requirements for the security of important data and personal information respectively, and AI developers must comply with relevant regulations when processing data. In particular, national security and public interest should be safeguarded when dealing with important data, and the rights and interests of natural persons should be protected when dealing with personal information. In 2020, the newly released *Civil Code of the People's Republic of China* (*Civil Code*) protects privacy and personal information interests. In 2021, the *Data Security Law* and *Personal Information Protection Law* (hereinafter referred to as 'PIPL') jointly provided for a more comprehensive approach to data governance. Responsible AI is ensured through new legal rules in four major dimensions in the field of data governance: giving individuals new civil rights, setting out obligations for processors, building a governance system for data security risk, and strengthening data processing responsibilities.

The new civil rights granted to individuals are mainly reflected in Chapter 4 of the *PIPL*. Also, *Civil Code* has laid down a 'privacy right' and a 'personal information right.' Privacy refers to a natural person's undisturbed private life and the private space, private activities, and private information that the person does not want others to know about, while personal information is recorded electronically or by other means that can be used, by itself or in combination with other information, to identify a natural person.[19] Such distinctions may not be marked and are rarely mentioned in legal and academic research in Europe and the United States (US). However, through these two systems, China constructs the strict protection of privacy rights, protecting natural persons from being exposed or interfered with and giving them the right to keep personal information from being handled illegally. According to the *Civil Code*, the private information included in personal information shall apply to the provisions of privacy; if there is no such provision, the provisions on the protection of personal information, such as the *PIPL*, shall be applied. The *PIPL* provides a series of specific rights in Articles 44–55, the content of which is consistent with the connotation of some articles of the European Union (EU) *General*

[17] Weixing Shen and Yun Liu, 'New Paradigm of Legal Research: Connotation, Category and Method of Computational Law' (2020) 5 *Chinese Journal of Law* 3–23.

[18] Notice of the State Council on Printing and Distributing the Action Platform for Promoting the Development of Big Data, Document No GF [2015] No 50, issued by the State Council on 31 August 2015.

[19] Article 1032 and Article 1034 of the Civil Code of the People's Republic of China (Adopted at the Third Session of the Thirteenth National People's Congress on 28 May 2020), Order No 45 of the President of the People's Republic of China (hereafter Civil Code of the People's Republic of China).

TABLE 9.1. *Individuals' Rights in Personal Information Processing Activities*

No.	Name of right	Legal references
1	The right to be informed, to decide, to restrict or refuse the processing	PIPL Art. 44
2	The right to consult, duplicate and transfer personal information	PIPL Art. 45
3	The right of correction or supplementation of their personal information	PIPL Art. 46
4	The right to delete	PIPL Art. 47
5	The right to request personal information processors to explain their personal information processing rules.	PIPL Art. 48
6	Exercise the rights of the relevant personal information of the deceased	PIPL Art. 49
7	The right to get remedy	PIPL Art. 50
8	Privacy respected	CC Art. 1032

Note: PIPL refers to *Personal Information Protection Law of the People's Republic of China*; CC refers to *Civil Code of the People's Republic of China*; DSL refers to *Data Security Law of the People's Republic of China*.

TABLE 9.2. *Obligations of Data Processors*

No.	Name of obligation	Legal reference
1	Acquire legal basis for processing personal information	PIPL Art. 13
2	Truthfully, accurately, and completely notify individuals of the relevant matters in a conspicuous way and in clear and easily understood language	PIPL Art. 14 & 17
3	Take corresponding security technical measures	PIPL Art. 51 & 59
4	Appoint a person in charge of personal information protection	PIPL Art. 52 & 53
5	Audit on a regular basis the compliance of their processing of personal information	PIPL Art. 54
6	Conduct personal information protection impact assessment in advance, and record the processing information	PIPL Art. 55 & 56
7	Immediately take remedial measures, and notify the authority performing personal information protection functions and the relevant individuals	PIPL Art. 57 DSL Art. 29 & 30
8	Specific requirements for sharing data	PIPL Art. 23
9	Specific requirements for important Internet platforms	PIPL Art. 58

Note: PIPL refers to *Personal Information Protection Law of the People's Republic of China*; CC refers to *Civil Code of the People's Republic of China*; DSL refers to *Data Security Law of the People's Republic of China*.

Data Protection Regulation (GDPR),[20] the fundamental purpose of which is to safeguard the rights of individuals in the data processing environment. Based on the protection of these rights, when AI handles personal information, it is also necessary to fully respect human dignity and ensure that personal information is not plundered by information technology. See *Table 9.1* for the specific rights system and its legal basis.

The obligations of processors are set not only to protect the personal information rights and interests of natural persons but also to strengthen the regulatory measures of protection specifically. AI developers and operators may be personal information processors, and they are subjected to comply with the nine major obligations under the *PIPL* and the *Data Security Law*, as shown in *Table 9.2*. These obligations cover the entire life cycle of personal information processing,

[20] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

TABLE 9.3. *Risk Management System of Data Governance*

No.	Risk Management System	Legal reference
1	Informed consent	PIPL Art. 13–17
2	Data minimization	PIPL Art. 6 & 19
3	Openness and transparency	PIPL Art. 7, 17, 48, 58
4	Cross-border security management system	PIPL Art. 38–43 DSL Art. 31
5	Sensitive personal information processing rules	PIPL Art. 28–32
6	Categorized and hierarchical data protection system	DSL Art. 21
7	Risk monitoring and security emergency response and disposition mechanism	DSL Art. 22 & 23
8	Public supervising for personal information system	PIPL Art. 60–65 DSL Art. 40

Note: PIPL refers to *Personal Information Protection Law of the People's Republic of China*; DSL refers to *Data Security Law of the People's Republic of China*.

ensuring the accountability of AI applications and reducing or eliminating the risk of damage to personal information.

Once data security risks in AI applications arise, it is difficult to recover from the damage. In order to avoid different types of data security risks such as personal information and important data, *Data Security Law* and *PIPL* establish a series of mechanisms to identify, eliminate, and resolve risks, thus ensuring data security in AI applications. Risk governance measures can be understood from different dimensions. Eight important governance measures under the law are listed in *Table* 9.3.

One of the basic principles for responsible AI is accountability, which also applies to data governance. The developers, controllers, and operators of AI systems can also be regarded as personal information processors on *PIPL* or data processors on *Data Security Law*, and they must comply with the above obligations. If the relevant obligators of the AI system violate data security obligations, they are liable for the corresponding damage consequences. The liability includes civil liability for compensation, administrative penalty, and criminal liability. Chapter VI of the *Data Security Act* and Chapter VII of the PIPL provide a number of legal liabilities that can ensure that individuals are able to obtain remedies and processors are punished in the event of data risks.

2. *Responsible AI Based on Algorithm Governance*

Responsible AI requires a combination of external and internal factors to play an active role. Data is the external factor, and algorithm is the internal factor. Algorithms are the key components of intelligence, and a series of algorithms combined with data training can form an AI system. Here is an example of the intelligent trial system developed by the authors' research group in a research program on Intelligent Assistive Technology in the Context of Judicial Process Involving Concerned Civil and Commercial Cases for the China courts. The actual workflow of this platform can be represented in Figure 9.1.

This process is not a unique way to develop an AI system, but it is an example of common practice. Through the earlier mentioned program development process, a computational function can be implemented, that is, data input, model calculation, and data output, where the

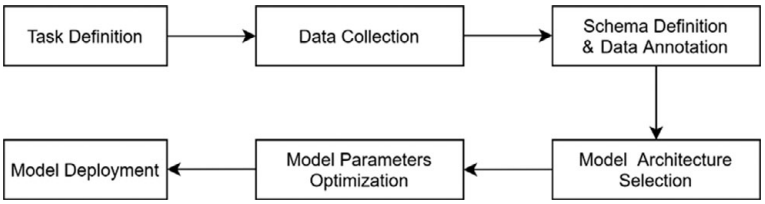

FIGURE 9.1 A development process of AI application

merit of the model directly determines the performance of the AI system. In this process, pre-trained models are often selected to reduce the workload of development. The algorithm used combined these pre-trained models, and the new model structure formed after development changed some of the parameters. Therefore, the algorithm governance commonly used in practice mainly refers to the parameters in the model structure, and this chapter continues to adopt 'algorithmic governance' as a unified concept to continue the academic terminology.

The *E-Commerce Law*, the *PIPL* and other related laws provide relevant provisions on algorithm governance. On 27 August 2021, the Cyberspace Administration of China released the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)*, which provides many algorithm governance requirements. In the *E-Commerce Law*, the representative concept in the law of algorithm governance can be summarized as 'personalized recommendation'.[21] In the *PIPL*, the representative concept in this law of algorithmic governance is 'automated decision-making'.[22] In the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)*, the representative concept of algorithm governance is 'algorithm recommendation technology' which refers to providing information by using algorithmic techniques such as generation synthesis, personalized push, sorting selection, retrieval and filtering, scheduling decision, etc.[23] Although the name of this regulation appears to apply to information services, information services here can be understood in a broad sense as information service technology.

We generally believe that the main principled requirements of algorithm governance are transparency, fairness, controllability, and accountability. The governance of algorithms in relevant Chinese laws and regulations basically follows the above-mentioned principles, which are also reflected in this regulation. According to the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)*, the actions to use algorithm recommendations should follow the principles of fairness, openness, transparency, reasonableness, and honesty.[24] Moreover, it explicitly prohibits the use of algorithm recommendation services to engage in activities prohibited by laws and regulations, such as endangering national security, disrupting the economic and social order, and infringing on the legitimate rights and interests of others.[25]

[21] Article 18 of the E-Commerce Law.
[22] Article 73 of the Personal Information Protection Law (effective 1 November 2021).
[23] Article 2 of the Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).
[24] Article 4 of the Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).
[25] Article 6 of the Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).

Data-driven AI has a certain degree of incomprehensibility, and algorithmic transparency can help us understand how AI systems work and ensure that users make well-informed choices about their use behavior. According to the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)*, the algorithm recommendation service provider should inform users of the algorithm recommendation service in a conspicuous manner and publicize the basic principle, purpose, and operation mechanism of the algorithm recommendation service properly.[26] In addition, Articles 24 and 48 of *PIPL* also have the requirement of algorithm transparency, which makes it clear that individuals have the right to request the personal information processors to explain their personal information processing rules. Individuals have the right to request the personal information processors to explain the decisions that significantly impact their rights and interests through automated decision-making, and the right to refuse to allow the personal information processors to make decisions through automated decision-making only.

Algorithm bias is also a highly controversial issue in algorithm governance, and the center topic is how to ensure the fairness of AI. In China, establishing higher prices for price-insensitive users through algorithms occasionally occurs in e-commerce. The main scenario occurs when cheaper prices are set for new users while relatively higher prices are set for older users who have developed a dependency. Article 18 of the *E-Commerce Law* and Article 21 of the *PIPL* have already made relevant provisions. Articles 10 and 18 of the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)* have made further provisions: providers of algorithm recommendation service shall strengthen the management of user models and user labels and, to improve the rules of interest points recorded in the user model, they shall not record illegal keywords or unhealthy information in the user's interest points, or mark it as user labels to recommend information. And they shall not set discriminatory or prejudicial user labels. Algorithm recommendation service providers selling goods or providing services to consumers shall protect consumers' legitimate rights and interests and shall not use algorithms to impose unreasonable differential treatment in transaction prices and other transaction prices based on consumer preferences, transaction conditions, and other characteristics or illegal acts.

At present, Chinese e-commerce operators still have different opinions on whether such behavior constitutes algorithmic bias. However, with extensive news media coverage, the general public opinion is more inclined to oppose algorithmic biases such as differential treatment of older users.

AI replaces some human behavior with automatic machine behaviors, and controllability is the essential requirement to ensure the safety and stability of AI. In order to prevent the risk of loss of control, the *Regulations on the Promotion of Artificial Intelligence Industry in the Shenzhen Special Economic Zone (Draft)* was released in July 2021. It set out the rules for agile governance, that is, organizing and conducting social experiments on AI; studying the comprehensive influence of AI development on the behavior patterns, social psychology, employment structure, income changes, social equity of individuals and organizations; and accumulating data and practical experience.[27]

[26] Article 14 of the Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).

[27] Article 66 of the Regulations of Shenzhen Special Economic Zone on the Promotion of Artificial Intelligence Industry (Draft for Soliciting Public Comment) (14 July 2021).

At present, China mainly focuses on self-driving cars and so-called robot advisors that give advice with regard to investment decisions. Relevant departments of the State Council and some localities have issued a series of road-testing specifications for intelligent connected vehicles (ICV), making closed road testing a prerequisite for self-driving cars to be put on the market. At the same time, to further improve their controllability, the cars tested on designated open roads must also be equipped with drivers ready to take over.[28]

On the other hand, the *Guidance on Standardizing the Asset Management Business of Financial Institutions* issued by the People's Bank of China and other departments in 2018 also points out the requirements of uncontrolled risk prevention in the field of smart investment consultants. It states that

> Financial institutions should develop corresponding AI algorithms or program trading according to different product investment strategies to avoid algorithm homogeneity and increase procyclicality of investment behavior, and for the resulting market volatility risk to develop a response plan. Due to algorithm homogenization, programming design errors, insufficient depth of data utilization and other Artificial Intelligence algorithm model defects or system anomalies, resulting in herding effects, affecting the stable operation of financial markets, financial institutions should promptly take manual intervention measures to force the adjustment or termination of the Artificial Intelligence business.[29]

The accountability of algorithms requires that regulators and stakeholders perform their respective duties to ensure that technological innovation is accompanied by effective risk mitigation. In terms of China's legal system, the *Civil Code*, the *Product Quality Law*, and other related laws provide the basis for the accountability of the algorithm.

For example, the *Product Quality Law* requires producers who design and sell products to reach the best (not the highest) degree of care; at the same time, it imposes strict liability control over unreasonable risks and aims that producers of AI system products improve their controllability requirements. In the *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment)*, the Cyberspace Administration of China expected a new rule that providers of algorithm recommendation services with high risk should register within ten working days from the date of service. The information of the service provider name, service form, application domain, algorithm type, algorithm self-evaluation report, and content to be publicized should be provided through the algorithm filing system of Internet information service.[30] On the other hand, the service providers of algorithm recommendation should accept social supervision, set up a convenient complaint reporting portal, and handle public complaints and reports promptly. The algorithm recommendation service provider should establish a use complaint channel and system, standardize handling complaints and feedback in a timely fashion, and protect users' legitimate rights and interests.[31]

[28] The Ministry of Industry and Information Technology, the Ministry of Public Security and the Ministry of Transport, 'Specifications for Road Test Management of Intelligent Networked Vehicles (for Trial Implementation)' (3 April 2018) and The Ministry of Industry and Information Technology, the Ministry of Public Security and the Ministry of Transport, 'Regulations on the Management of Intelligent Networked Vehicles in Shenzhen Special Economic Zone (Draft for Soliciting Public Comment)' (23 March 2021).

[29] Article 23 of Guiding Opinions on Regulating Asset Management Business of Financial Institutions (27 April 2018, revised 31 July 2020) No 16 [2018] People's Bank of China (hereafter Guiding Opinions on Regulating Asset Management).

[30] Article 20 of the Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).

[31] Article 26 of the Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).

The content discussed above reflects the need for administrative authorities, algorithm developers and other relevant parties to fulfill their corresponding responsibilities under the accountability requirements of algorithms.

3. *Responsible AI Based on Platform Governance*

The world's primary AI technology innovation enterprises are companies of online platforms. These online platforms have strong technological innovation capabilities and a wide range of AI application scenarios, and many new technologies and new business models derive from online platforms. Therefore, platform governance is also an important aspect of achieving responsible AI. In China, online platform governance is mainly regulated in the *E-Commerce Law*, competition law, and other relevant laws and regulations. In recent years, provisions on AI governance under online platforms have been adopted or promulgated through the process of legislation or amendment.

There are a growing number of platforms in online transactions that use AI to determine flexible transaction rules. The *E-Commerce Law* issued in 2018 explicitly defines the platform as a regulated object, which requires that e-commerce platform operators should follow the principles of openness, fairness, and impartiality in formulating platform service agreements and transaction rules.[32] Article 18 *E-Commerce Law* also requires e-commerce operators to respect and equally protect the legitimate rights and interests of consumers when providing personalized recommendation services. In addition, the *Interim Provisions on the Management of Online Tourism Operation Services* issued in August 2020 also sets out that online travel operators shall not abuse technical means such as big data analysis to set unfair trading conditions based on tourists' consumption records and tourism preferences, to infringe on the legitimate rights and interests of tourists.[33] Thus, it is clear that the *E-Commerce Law* mainly requires that the application of AI should not undermine the rights of consumers and operators within the platform to be treated fairly.

Online platforms often use AI to gain an unfair market competitive advantage. China's *Anti-Monopoly Law*, *Anti-Unfair Competition Law*, and other relative regulations are also concerned with the platform responsibilities in the process of AI application.[34] In terms of horizontal monopoly agreements, the substantial existence of coordination through data, algorithms, platform rules, or other means is regarded as an illegal monopoly. In terms of vertical monopoly agreements, it is also regarded as an illegal monopoly to exclude or restrict market competition through directly or indirectly limiting the price by using data and algorithms, or limiting other transaction conditions by using technical means, platform rules, data, and algorithms.[35] The use of big data and algorithms to impose differential prices or other trading conditions or to impose differentiated standards, rules, or algorithms based on the ability to pay,[36] consumption preferences, and usage habits of the counterparty is also considered an illegal monopolistic act of abuse of a dominant position in the market. AI used to implement these monopolistic acts mainly refers to large-scale Internet platforms with a dominant market position.

[32] Article 32 of the E-Commerce Law.
[33] Article 15 of the Interim Provisions on the Management of Online Tourism Operation Services, Order No 4 of the Ministry of Culture and Tourism of the People's Republic of China (2020).
[34] Article 5 of the Anti-Monopoly Guidelines on Platform Economy.
[35] Article 7 of the Anti-Monopoly Guidelines on Platform Economy.
[36] Article 17 of the Anti-Monopoly Guidelines on Platform Economy.

It is an act of unfair competition for a business operator to use data, algorithms and other technical means to implement traffic hijacking, interference and malicious incompatibility to prevent or disrupt the normal operation of network products or services lawfully provided by other operators.[37] Similarly, operators that use data, algorithms, and other technical means to unreasonably provide different transaction information to counterparties managed by the same transaction conditions (by collecting and analysing transaction information, the content, and time of internet browsing; the brand and value of the terminal equipment used for the transaction; etc.), are infringing the counterparties' right to know, right to choose, right to fair trade, etc., and are disrupting the fair trading order of the market[38] Those who use AI to implement the above unfair competition can be both large Internet platforms and participants of other platform markets.

4. *Responsible AI under Specific Scenarios*

Specific scenarios in the field of AI based on new technologies and new business models often attract special attention. As a result, some AI-related regulations in different areas have been emerging. For example, China has special regulations to ensure responsible AI in areas such as labor employment, facial recognition, autonomous driving, smart investment consultants, deep forgery, online travel, online litigation, etc. The regulations related to responsible AI in these special areas can be divided into two categories. The first category is to reflect the provisions of the *PIPL*, the *E-Commerce Law*, and other relevant regulations to these specific areas, which increases the relevance of norm implementation but does not substantially introduce a new legal system. The second category is to establish more legal obligations and rights based on the special circumstances in the specific scenarios.

In the labor market, in September 2020, a widely spread report in the social platforms of China about the abuse of algorithms for performance management on online platforms revealed that a series of automated behaviors based on algorithms, such as point rating systems, system 'upgrades' to shorten delivery times, and navigation instructions that violate traffic rules, are forcing couriers to engage in high-intensity labor.[39] In July 2021, the State Administration for Market Regulation (SAMR) and relevant departments jointly issued binding opinions pointing out that the network catering platform and third-party partners should reasonably set the performance appraisal system for delivery workers. In the development of adjustments to the assessment, rewards and punishments and other systems or significant matters involving the delivery workers' direct interests should be publicized in advance to fully listen to the views of the delivery workers, trade unions, and other parties. Optimize the algorithm rules, not the 'strictest algorithm' as the assessment requirements, through the 'algorithm to take' and other ways to reasonably determine the number of orders, online rate, and other assessment factors, to determine appropriate flexibility of the delivery time frame.[40] The *Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting*

[37] Article 13 of the Notice from the State Administration for Market Regulation of the Provisions on Prohibited Acts of Unfair Competition Online (Draft for Soliciting Public Comment).

[38] Article 21 of the Notice from the State Administration for Market Regulation of the Provisions on Prohibited Acts of Unfair Competition Online (Draft for Soliciting Public Comment).

[39] Lai Youxuan, 'Deliveries, Stuck in the System' (*People*, September 2020) https://epaper.gmw.cn/wzb/html/2020-09/12/nw.D110000wzb_20200912_1-01.htm.

[40] Guidance on the Implementation of The Responsibility of Online Catering Platforms to Effectively Safeguard the Rights and Interests of Food Delivery Workers, issued by SAMR on 16 July 2021.

Public Comment) released in August 2021 further points out that the algorithm recommendation service providers providing work scheduling services to workers should establish and improve the platform order distribution, compensation and payment, working hours, rewards and punishments and other related algorithms, and fulfill the obligations of workers' rights and interests.[41] These requirements are a special case in the field of labor employment and reflect the principle of inclusiveness of AI to avoid the risk of polarization of AI.

In the field of facial recognition, an associate professor of the law school of Zhejiang University of Technology sued for the compulsory use of facial recognition equipment as admission to Hangzhou Safari Park, which was regarded as the first case of facial recognition rights defense in the judicial field in China. After that, a professor of the law school of Tsinghua University published a criticism for the compulsory use of facial recognition equipment as permission to enter into the apartment. A series of facial recognition incidents have raised social concerns about the right to choose facial recognition applications. In December 2020, the Cyberspace Administration of China drafted a *Security Management Regulations for Commercial Applications of Facial Recognition Technology (Draft)*, but the draft has not yet been released. Meanwhile, the Supreme People's Court has published a judicial opinion, *Provisions on Several Issues Concerning the Application of Law in Hearing Civil Cases Relating to the Use of Facial Recognition Technology for Handling Personal Information*, which states that: if building managers use facial recognition systems as the only way to verify owners or property users (to enter or leave the building), the people's court shall support the owners or property users who disagree with using facial recognition and request to provide other reasonable verification methods under the law.[42] In addition, the Legal Working Committee of the Standing Committee of the National People's Congress is also considering drafting special legal provisions for facial recognition. These moves reflect the requirements of personal biometric information protection and combine the specific scenarios of facial recognition to make special rules.

Furthermore, in the field of autonomous driving, the relevant departments of the State Council have specifically set complex conditions for its testing on open roads, including commissioned inspection reports of autonomous driving functions issued by third-party testing agencies, testing programs, and certificates of compulsory insurance for traffic accidents.[43] In the field of smart investment consultants, financial institutions should report the main parameters of the AI model and the main logic of asset allocation to the financial supervision and management authorities, set up separate smart management accounts for investors, fully indicate the inherent flaws and risks of using AI algorithms, clarify the transaction process, strengthen the management of the traces, and strictly monitor the trading positions, risk limits, types of transactions, pricing authority of smart management accounts, etc. Financial institutions that cause losses to investors due to violations of law or mismanagement shall be liable for damages as prescribed by law.[44] In the field of deep-fakes governance, the law requires that no organization or individual shall infringe upon the portrait rights of others by scandalizing, defacing, or using information technology means to forge, etc.; for the protection of the voice of natural persons, the relevant provisions on the protection of portrait rights shall apply by reference.[45] The dispersion of the above provisions indicates that there are

[41] Article 17 of Regulation on Internet Information Service Based on Algorithm Recommendation Technology (Draft for Soliciting Public Comment).
[42] Article 10 of Supreme People's Court, Provisions on Facial Recognition (n 16).
[43] Article 9 of the Intelligent Networked Vehicle Road Test Management Specifications (for Trial Implementation) notice.
[44] Guiding Opinions on Regulating Asset Management (n 29).
[45] Article 1019 and Article 1023 of the Civil Code of the People's Republic of China (n 19).

differences in the degree of application of AI in different fields, and there are differences in the risks and security needs arising from its use in different fields. In the absence of comprehensive legislation on AI, the adoption of special binding provisions or guidance to address specific issues is a way to balance the pursuit of development with security values.

IV. CONCLUSION

The legal professional culture is generally conservative, which results in the law and regulations always lagging behind in responding to innovation and new technologies. At the beginning of the rapid development of AI, we have mainly implemented risk governance through moral codes, ethical guidelines, and technical standards. In contrast to these soft laws, the national legislature can enact mandatory 'hard laws' that establish general and binding rules on the scope of application, management system, safety measures, rights and remedies, and legal liabilities of AI technologies. China has issued several soft law governance tools around responsible AI in different sectors but does not yet have comprehensive AI legislation. However, China is still attempting to develop comprehensive AI legislation as evidenced by *President Xi Jinping's* statement on AI legislation, the requirements in the national-level development plan for a new generation of AI, the attention paid to the topic by the National People's Congress deputies, and some local legislative motions, as represented by Shenzhen.

The law has been out of date since its enactment. However, this does not mean that the law can do nothing about the problems after its promulgation. In the codified tradition, the applicability of various legal documents is often scalable, which enables new technologies and new business models to find corresponding applicable provisions. The effective *Civil Code*, *E-Commerce Law*, *Product Quality Law*, and other relevant legislations can serve as legal requirements developing responsible AI. In addition, new laws and other binding documents enacted in recent years provide a substantial basis for AI governance, and the effective and draft documents released in 2021 show that responsible AI is increasingly a concrete goal that needs to be enforced. Looking toward the future, two different options for legislative routes exist in countries including China, the EU, and the United States. One option is a foresighted legal design mindset that designs an institutional track to develop emerging technologies in the fastest possible way. Under this option, after the basic application pattern of AI technology is formed, lawmakers will summarize and predict the various risks of AI based on the existing situation and the understanding obtained by reasoning. Another option is to adopt a 'wait and see' approach, arguing that it is still too early for lawmakers to see just how this technology will impact citizens. Under this option, lawmakers pay more attention to the positive value of emerging technology development, and the risks involved are self-identified, self-adjusted, self-regulated, and self-healed by the free competition mechanism of the market.

From the current legislative dynamics in China, improving regulations related to data, algorithms, platforms, and specific scenarios will provide a broad and effective basis for AI governance. The development of comprehensive AI legislation has not been formally included in the NPC Standing Committee's work plan in the short term, but this does not prevent local legislatures from exploring the possibility of comprehensive legislation. If a comprehensive AI legislation is to be enacted, its key elements are to record the types of AI risks, design the mechanism for identifying AI risks, and construct the mechanism for resolving AI risks. The EU proposal of AI Act released in 2021 has also been widely followed in China, and we can expect that after the proposal of the act is passed in Europe, it is likely that similar legislation will be enacted in China shortly thereafter.

10

Towards a Global Artificial Intelligence Charter

*Thomas Metzinger**

I. INTRODUCTION

The time has come to move the ongoing public debate on Artificial Intelligence (AI) into our political institutions. Many experts believe that during the next decade we will be confronted with an inflection point in history and that there is a closing window of opportunity for working out the applied ethics of AI. Political institutions must, therefore, produce and implement a minimal but sufficient set of ethical and legal constraints for the beneficial use and future development of AI. They must also create a rational, evidence-based process of critical discussion aimed at continuously updating, improving, and revising this first set of normative constraints. Given the current situation, the default outcome is that the values guiding AI development will be set by a very small number of human beings acting within large private corporations and military institutions. Therefore, one goal is to proactively integrate as many perspectives as possible – and in a timely manner. Many initiatives have already sprung up worldwide and are actively investigating recent advances in AI in relation to issues concerning applied ethics, including its legal aspects, future sociocultural implications, existential risks, and policymaking.[1] Public debate is heated, and some may even have the impression that major

* This is an updated and considerably expanded version of a chapter that goes back to a lecture I gave on 19 October 2017 at the European Parliament in Brussels (Belgium). Cf. (2018), Towards a Global Artificial Intelligence Charter. In European Parliament (ed), Should we fear artificial intelligence? PE 614.547. www.philosophie.fb05.uni-mainz.de/files/2018/03/Metzinger_2018_Global_Artificial_Intelligence_Charter_PE_614.547.pdf.

[1] For an overview of existing initiatives, I recommend T Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines' (2020) 30 *Minds & Machines* 99 https://doi.org/10.1007/s11023-020-09517-8; and the AI Ethics Guidelines Global Inventory created by Algorithm Watch, at https://inventory.algorithmwatch.org/. Other helpful overviews are S Baum, 'A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy' (2017) Global Catastrophic Risk Institute Working Paper 17-1 https://ssrn.com/abstract=3070741; P Boddington, *Towards a Code of Ethics for Artificial Intelligence* (2017) 3. I have refrained from providing full documentation here, but useful entry points into the literature are A Mannino and others, 'Artificial Intelligence. Opportunities and Risks' (2015) 2 Policy Papers of the Effective Altruism Foundation https://ea-foundation.org/files/ai-opportunities-and-risks.pdf (hereafter Mannino et al., 'Opportunities and Risks'); P Stone and others, 'Artificial Intelligence and Life in 2030' (2016) One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel https://ai100.stanford.edu/2016-report; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 'Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems' (*IEEE*, 2017) http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html; N Bostrom, A Dafoe, and C Flynn, 'Policy Desiderata in the Development of Machine Superintelligence' (2017) Oxford University Working Paper www.nickbostrom.com/papers/aipolicy.pdf; M Madary and T Metzinger, 'Real Virtuality. A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology' (2016) 3 *Frontiers in Robotics and AI* 3 http://journal.frontiersin.org/article/10.3389/frobt.2016.00003/full.

political institutions like the European Union (EU) are unable to react with adequate speed to new technological risks and to rising concern amongst the general public. We should, therefore, increase the agility, efficiency, and systematicity of current political efforts to implement rules by developing a more formal and institutionalised democratic process, and perhaps even new models of governance.

To initiate a more systematic and structured process, I will present a concise and non-exclusive list of the five most important problem domains, each with practical recommendations. The first problem domain to be examined is the one that, in my view, is made up of those issues that have the smallest chance of being solved. It should, therefore, be approached in a multilayered process, beginning in the EU itself.

II. THE 'RACE-TO-THE-BOTTOM' PROBLEM

We need to develop and implement worldwide safety standards for AI research. A Global Charter for AI is necessary, because such safety standards can be effective only if they involve a binding commitment to certain rules by all countries participating and investing in the relevant type of research and development. Given the current competitive economic and military context, the safety of AI research will very likely be reduced in favour of more rapid progress and reduced cost, namely by moving it to countries with low safety standards and low political transparency (an obvious and strong analogy is the problem of tax evasion by corporations and trusts). If international cooperation and coordination succeeded, then a 'race to the bottom' in safety standards (through the relocation of scientific and industrial AI research) could, in principle, be avoided. However, the current landscape of incentives makes this a highly unlikely outcome. Non-democratic political actors, financiers, and industrial lobbyists will almost certainly prevent any more serious globalised approach to AI ethics.[2] I think that, for most of the goals I will sketch below, it would not be intellectually honest to assume that they can actually be realised, at least not in any realistic time frame and with the necessary speed (this is particularly true of Recommendations 2, 4, 6, 7, 10, 12, and 14). Nevertheless, it may be helpful to formulate a general set of desiderata to help structure future debates.

Recommendation 1
The EU should immediately develop a European AI Charter.

Recommendation 2
In parallel, the EU should initiate a political process steering the development of a Global AI Charter.

Recommendation 3
The EU should invest resources into systematic strengthening of international cooperation and coordination. Strategic mistrust should be minimised; commonalities can be defined via maximally negative scenarios.

[2] T Metzinger, 'Ethics Washing Made in Europe' *Tagesspiegel* (8 April 2019) www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

The second problem domain to be examined is arguably constituted by the most urgent set of issues, and these also have a fairly small chance of being adequately resolved.

III. PREVENTION OF AN AI ARMS RACE

It is in the interests of the citizens of the EU that an AI arms race, for example between China and the United States (US), be halted before it gathers real momentum. Again, it may well be too late for this, and European influence is obviously limited. However, research into, and development of, offensive autonomous weapons should not be funded, and indeed should be outright banned, on EU territory. Autonomous weapons select and engage targets without human intervention, and they will act and react on ever shorter timescales, which in turn will make it seem reasonable to transfer more and more human autonomy into these systems themselves. They may, therefore, create military contexts in which relinquishing human control almost entirely seems like the rational choice. Autonomous weapon systems lower the threshold for entering a war, and if both warring parties possess intelligent, autonomous weapon systems there is an increased danger of fast escalation based exclusively on machine-made decisions. In this problem domain, the degree of complexity is even higher than in the context of preventing the development and proliferation of nuclear weapons, for example, because most of the relevant research does not take place in public universities. In addition, if humanity forces itself into an arms race on this new technological level, the historical process of an arms race itself may become autonomous and resist political interventions.

Recommendation 4
The EU should ban all research on offensive autonomous weapons on its territory and seek international agreements on such prohibitions.

Recommendation 5
For purely defensive military applications (if they are at all conceivable), the EU should fund research into the maximal degree of autonomy for intelligent systems that appears to be acceptable from an ethical and legal perspective.

Recommendation 6
On an international level, the EU should start a major initiative to prevent the emergence of an AI arms race, using all diplomatic and political instruments available.

The third problem domain to be examined is the one for which the predictive horizon is probably still quite distant, but where epistemic uncertainty is high and potential damage could be extremely large.

IV. A MORATORIUM ON SYNTHETIC PHENOMENOLOGY

It is important that all politicians understand the difference between AI and artificial consciousness. The unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. 'Synthetic phenomenology' (SP, a

term coined in analogy to 'synthetic biology') refers to the possibility of creating not only general intelligence, but also consciousness or subjective experiences, in advanced artificial systems. Potential future artificial subjects of experience currently have no representation in the current political process, they have no legal status, and their interests are not represented in any ethics committee. To make ethical decisions, it is important to have an understanding of which natural and artificial systems have the capacity for producing consciousness, and in particular for experiencing negative states like suffering.[3] One potential risk is that of dramatically increasing the overall amount of suffering in the universe, for example via cascades of copies or the rapid duplication of conscious systems on a vast scale.

For this, I refer readers to an open-access publication of mine, titled 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology'.[4] The risk that has to be minimised in a rational and evidence-based manner is the risk of an 'explosion of negative phenomenology' (ENP; or simply a 'suffering explosion') in advanced AI and other post-biotic systems. I will here define 'negative phenomenology' as any kind of conscious experience a conscious system would avoid or rather not go through if it had a choice.

On ethical grounds, we should not risk an explosion of conscious suffering – at the very least not before we have a much deeper scientific and philosophical understanding of what both consciousness and suffering really are. As we presently have no good theory of consciousness and no good, hardware-independent theory about what 'suffering' really is, the ENP risk is currently incalculable. It is unethical to run incalculable risks of this magnitude. Therefore, until 2050, there should be a global ban on all research that directly aims at, or indirectly and knowingly risks, the emergence of synthetic phenomenology.

Synthetic phenomenology is only one example of a type of risk to which political institutions have turned out to be systematically blind, typically dismissing such risks as 'mere science fiction'. It is equally important that all politicians understand both the possible interactions amongst specific risks and – given the large number of 'unknown unknowns' in this domain – the fact that there is an ethics of risk-taking itself. This point relates to uncomprehended risks we currently label as 'mid-term', 'long-term', or 'epistemically indeterminate'.

Recommendation 7

The EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory, and seek international agreements on such prohibitions.[5]

Recommendation 8

Given the current level of uncertainty and disagreement within the nascent field of machine consciousness, there is a pressing need to promote, fund, and coordinate relevant interdisciplinary research projects (comprising fields such as philosophy, neuroscience, and computer

[3] See T Metzinger, 'Two Principles for Robot Ethics' (2013) in E Hilgendorf and JP Günther (eds), *Robotik und Gesetzgebung*; T Metzinger, 'Suffering' (2017) in K Almqvist and A Haag (eds), *The Return of Consciousness*.

[4] See T Metzinger, 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology' (2021) 8 (1) *Journal of Artificial Intelligence and Consciousness* 1, 43–66. https://www.worldscientific.com/doi/pdf/10.1142/S270507852150003X.

[5] This includes approaches that aim at a confluence of neuroscience and AI with the specific aim of fostering the development of machine consciousness. For recent examples see S Dehaene, H Lau, and S Kouider, 'What Is Consciousness, and Could Machines Have It?' (2017) 6362 *Science* 486; MSA Graziano, 'The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness' (2017) 4 *Frontiers in Robotics and AI*; R Kanai, 'We Need Conscious Robots. How Introspection and Imagination Make Robots Better' (*Nautilus*, 27 April 2017) http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

science). Specific topics of relevance are evidence-based conceptual, neurobiological, and computational models of conscious experience, self-awareness, and suffering.

Recommendation 9
On the level of foundational research there is a need to promote, fund, and coordinate systematic research into the applied ethics of non-biological systems capable of conscious experience, self-awareness, and subjectively experienced suffering.

The next general problem domain to be examined is the most complex, and likely contains the largest number of unexpected problems and 'unknown unknowns'.

V. DANGERS TO SOCIAL COHESION

Advanced AI technology will clearly provide many possibilities for optimising the political process itself, including novel opportunities for rational, value-based social engineering and more efficient, evidence-based forms of governance. On the other hand, it is plausible to assume that there are many new, at present unknown, risks with the potential to undermine efforts to sustain social cohesion. It is also reasonable to assume the existence of a larger number of 'unknown unknowns', of AI-related risks that we will discover only by accident and late in the day. Therefore, the EU should allocate separate resources to prepare for situations in which such unexpected 'unknown unknowns' are suddenly discovered.

Many experts believe that the most proximal and well-defined risk is massive unemployment through automation.[6] The implementation of AI technology by financially potent stakeholders may lead to a steeper income gradient, increased inequality, and dangerous patterns of social stratification.[7] Concrete risks are extensive wage cuts, a collapse of income tax, plus an overload of social security systems. But AI poses many other risks for social cohesion, for example via privately owned and autonomously controlled social media aimed at harvesting human attention and 'packaging' it for further use by their customers, or in 'engineering' the formation of political will via Big Nudging strategies and AI-controlled choice architectures that are not transparent to the individual citizens whose behaviour is thus controlled.[8] Future AI technology will be extremely good at modelling and predictively controlling human behavior – for example by positive reinforcement and indirect suggestions, making compliance with certain norms or the emergence of 'motives' and decision outcomes appear entirely spontaneous and unforced. In combination with Big Nudging and predictive user control, intelligent surveillance technology could also increase global risks by locally helping to stabilise authoritarian regimes in an efficient manner. Again, most of these risks to social cohesion are still very likely unknown at present, and we may discover them only by accident. Policymakers must also understand that any technology that can purposefully optimise the intelligibility of its own action for human users can in principle also optimise for deception. Great care must therefore be taken to avoid accidental or even intended specification of the reward function of any AI in a way that might indirectly damage the common good.

[6] See European Parliamentary Research Service 'The Ethics of Artificial Intelligence: Issues and Initiatives' (*European Parliamentary Research Service*, 2020) 6–11.
[7] A Smith and J Anderson, 'AI, Robotics, and the Future of Jobs' (*Pew Research Center*, 2014) www.pewresearch.org/internet/wp-content/uploads/sites/9/2014/08/Future-of-AI-Robotics-and-Jobs.pdf.
[8] For a first set of references, see www.humanetech.com/brain-science.

AI technology is currently a private good. It is the duty of democratic political institutions to turn large portions of it into a well-protected common good, something that belongs to all of humanity. In the tragedy of the commons, everyone can often see what is coming, but if mechanisms for effectively counteracting the tragedy are not in existence it will unfold invisibly, for example in decentralised situations. The EU should proactively develop such preventative mechanisms.

Recommendation 10
Within the EU, AI-related productivity gains must be distributed in a socially just manner. Obviously, past practice and global trends clearly point in the opposite direction: We have (almost) never done this in the past, and existing financial incentives directly counteract this recommendation.

Recommendation 11
The EU should carefully research the potential for an unconditional basic income or a negative income tax on its territory.

Recommendation 12
Research programs are needed to investigate the feasibility of accurately timed initiatives for retraining threatened population strata towards creative and social skills.

The next problem domain is difficult to tackle because most of the cutting-edge research in AI has already moved out of publicly funded universities and research institutions. It is in the hands of private corporations, and, therefore, systematically non-transparent.

VI. RESEARCH ETHICS

One of the most difficult theoretical problems in this area is the problem of defining the conditions under which it would be rational to relinquish specific AI research pathways altogether (for instance, those involving the emergence of synthetic phenomenology, or plausibly engendering an explosive evolution of autonomously self-optimising systems not reliably aligned with human values). What would be concrete, minimal scenarios justifying a moratorium on certain branches of research? How will democratic institutions deal with deliberately unethical actors in a situation where collective decision-making is unrealistic and graded, and non-global forms of *ad hoc* cooperation have to be created? Similar issues have already occurred in so-called gain-of-function research involving experimentation aiming at an increase in the transmissibility and/or virulence of pathogens, such as certain highly pathogenic H5N1 influenza virus strains, smallpox, or anthrax. Here, influenza researchers laudably imposed a voluntary and temporary moratorium on themselves.[9] In principle, this could happen in the AI research community as well. Therefore, the EU should certainly complement its AI charter with a concrete code of ethical conduct for researchers working in funded projects. However, the deeper goal would be to develop a more comprehensive culture of moral sensitivity within the relevant research communities themselves. Rational, evidence-based identification and

[9] See FS Collins and AS Fauci, 'NIH Statement on H5N1' (*The NIH Director*, 2012) www.nih.gov/about-nih/who-we-are/nih-director/statements/nih-statement-h5n1; and RAM Fouchier and others, 'Pause on Avian Flu Transmission Studies' (2012) *Nature* 443.

minimisation of risks (including those pertaining to a distant future) ought to be a part of research itself, and scientists should cultivate a proactive attitude to risk, especially if they are likely to be the first to become aware of novel types of risk through their own work. Communication with the public, if needed, should be self-initiated, in the spirit of taking control and acting in advance of a possible future situation, rather than just reacting to criticism by non-experts with some set of pre-existing, formal rules. As *Michael Madary* and I note in our ethical code of conduct for virtual reality, which includes recommendations for good scientific practice: 'Scientists must understand that following a code of ethics is not the same as being ethical. A domain-specific ethics code, however consistent, developed and fine-grained future versions of it may be, can never function as a substitute for ethical reasoning itself.'[10]

Recommendation 13
Any AI Global Charter, or its European precursor, should always be complemented by a concrete Code of Ethical Conduct guiding researchers in their practical day-to-day work.

Recommendation 14
A new generation of applied ethicists specialised in problems of AI technology, autonomous systems, and related fields needs to be trained. The EU should systematically and immediately invest in developing the future expertise needed within the relevant political institutions, and it should do so aiming at an above-average level of academic excellence and professionality.

VII. META-GOVERNANCE AND THE PACING GAP

As briefly pointed out in the introductory paragraph, the accelerating development of AI has perhaps become the paradigmatic example of an extreme mismatch between existing governmental approaches and what would be needed to optimise the risk/benefit ratio in a timely fashion. The growth of AI exemplifies how powerfully time pressure can constrain rational and evidence-based identification, assessment, and management of emerging risks; creation of ethical guidelines; and implementation of an enforceable set of legal rules. There is a 'pacing problem': Existing governance structures are simply unable to respond to the challenge fast enough; political oversight has already fallen far behind technological evolution.[11]

I am drawing attention to the current situation not because I want to strike an alarmist tone or to end on a dystopian, pessimistic note. Rather, my point is that the adaptation of governance structures themselves is part of the problem landscape: In order to close or at least minimise the pacing gap, we have to invest resources into changing the structure of governance approaches

[10] M Madary and T Metzinger, 'Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology' (2016) 3(3) *Frontiers in Robotics and AI* 1, 12.

[11] GE Marchant, 'The Growing Gap between Emerging Technologies and the Law' in GE Marchant, BR Allenby, and JR Herkert (eds), *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight* (2011), 19, puts the general point very clearly in the abstract of a recent book chapter: 'Emerging technologies are developing at an ever-accelerating pace, whereas legal mechanisms for potential oversight are, if anything, slowing down. Legislation is often gridlocked, regulation is frequently ossified, and judicial proceedings are sometimes described as proceeding at a glacial pace. There are two consequences of this mismatch between the speeds of technology and law. First, some problems are overseen by regulatory frameworks that are increasingly obsolete and outdated. Second, other problems lack any meaningful oversight altogether. To address this growing gap between law and regulation, new legal tools, approaches, and mechanisms will be needed. Business as usual will not suffice'.

themselves. 'Meta-governance' means just this: A governance of governance equal to facing the risks and potential benefits of an explosive growth in specific sectors of technological development. For example, *Wendell Wallach* has pointed out that the effective oversight of emerging technologies requires some combination of both hard regulations enforced by government agencies and expanded soft-governance mechanisms.[12] *Gary Marchant* and *Wendell Wallach* have, therefore, proposed so-called Governance Coordination Committees (GCCs), a new type of institution providing a mechanism for coordinating and synchronising what they aptly describe as an 'explosion of governance strategies, actions, proposals, and institutions'[13] with existing work in established political institutions. A GCC for AI could act as an 'issue manager' for one specific, rapidly emerging technology; as an information clearinghouse, an early warning system, an analysis and monitoring instrument, and an international best-practice evaluator; and as an independent and trusted 'go-to' source for ethicists, media, scientists, and interested stakeholders. As *Marchant* and *Wallach* write: 'The influence of a GCC in meeting the critical need for a central coordinating entity will depend on its ability to establish itself as an honest broker that is respected by all relevant stakeholders.'[14]

Many other strategies and governance approaches are, of course, conceivable. However, this is not the place to discuss details. Here, the general point is simply that we can meet the challenge posed by rapid developments in AI and autonomous systems only if we put the question of meta-governance on top of our agenda right from the start. In Europe, the main obstacle to reaching this goal is, of course, 'soft corruption' through the Big Tech industrial lobby in Brussels: There are strong financial incentives and major actors involved in keeping the pacing gap as wide open as possible for as long as possible.[15]

Recommendation 15
The EU should invest in researching and developing new governance structures that dramatically increase the speed at which established political institutions can respond to problems and actually enforce new regulations.

VIII. CONCLUSION

I have proposed that the European Union immediately begin working towards the development of a Global AI Charter, in a multilayered process starting with an AI Charter for the EU itself. To briefly illustrate some of the core issues from my own perspective as a philosopher, I have identified five major thematic domains and provided 15 general recommendations for critical discussion. Obviously, this contribution was not meant as an exclusive or exhaustive list of the relevant issues. On the contrary, at its core, the applied ethics of AI is not a field for grand theories or ideological debates at all, but mostly a problem of sober, rational risk management involving different predictive horizons under great uncertainty. However, an important part of

[12] See W Wallach, *A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control* (2015), 250.
[13] This quote is taken from an unpublished, preliminary draft entitled 'An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics'; see also GE Marchant and W Wallach, 'Coordinating Technology Governance' (2015) 31 *Issues in Science and Technology* (hereafter Marchant and Wallach, 'Technology Governance').
[14] Marchant and Wallach, 'Technology Governance' (n 14), 47.
[15] For one recent report, see M Bank and others, 'Die Lobbymacht von Big Tech: Wie Google & Co die EU beeinflussen' (*Corporate Europe Observatory und LobbyControl e.V.*, 2021) www.lobbycontrol.de/wp-content/uploads/Studie_de_Lobbymacht-Big-Tech_31.8.21.pdf.

the problem is that we cannot rely on intuitions, because we must satisfy counterintuitive rationality constraints. Therefore, we also need humility, intellectual honesty, and genuine open-mindedness.

Let me end by quoting from a recent policy paper titled 'Artificial Intelligence: Opportunities and Risks', published by the Effective Altruism Foundation in Berlin, Germany:

> In decision situations where the stakes are very high, the following principles are of crucial importance:
>
> 1. Expensive precautions can be worth the cost even for low-probability risks, provided there is enough to win/lose thereby.
> 2. When there is little consensus in an area amongst experts, epistemic modesty is advisable. That is, one should not have too much confidence in the accuracy of one's own opinion either way.[16]

[16] Cf. Mannino and others, 'Opportunities and Risks' (n 1).

11

Intellectual Debt

With Great Power Comes Great Ignorance

*Jonathan Zittrain**

The boxes for prescription drugs typically include an insert of tissue-thin paper folded as tight as origami. For the bored or the preternaturally curious who unfurl it, there's a sketch of the drug's molecular structure using a notation that harkens to high school chemistry, along with 'Precautions' and 'Dosage and Administration' and 'How Supplied'. And for many drugs, under 'Clinical Pharmacology', one finds a sentence like this one for the wakefulness drug Provigil, after the subheading 'Mechanism of Action': 'The mechanism(s) through which modafinil promotes wakefulness is unknown.'[1] That sentence alone might provoke wakefulness without assistance from the drug. How is it that something could be so studied and scrutinized to find its way to regulatory approval and widespread prescribing, while we don't know how it works?

The answer is that industrial drug discovery has long taken the form of trial-and-error testing of new substances in, say, mice. If the creatures' condition is improved with no obvious downside, the drug may be suitable for human testing. Such a drug can then move through a trial process and earn approval. In some cases, its success might inspire new research to fill in the blanks on mechanism of action. For example, aspirin was discovered in 1897, and an explanation of how it works followed in 1995.[2] That, in turn, has spurred some research leads on making better pain relievers through something other than trial and error.

This kind of discovery – answers first, explanations later – accrues what I call 'intellectual debt'. We gain insight into what works without knowing why it works. We can put that insight to use immediately, and then tell ourselves we'll figure out the details later. Sometimes we pay off the debt quickly; sometimes, as with aspirin, it takes a century; and sometimes we never pay it off at all.

Be they of money or ideas, loans can offer great leverage. We can get the benefits of money – including use as investment to produce more wealth – before we've actually earned it, and we can deploy new ideas before having to plumb them to bedrock truth.

Indebtedness also carries risks. For intellectual debt, these risks can be quite profound, both because we are borrowing as a society, rather than individually, and because new technologies of Artificial Intelligence (AI) – specifically, machine learning – are bringing the old model of drug

* This chapter is based on an essay found at https://perma.cc/CN55-XLCW?type=image. A derivative version of it was published in *The New Yorker*, 'The Hidden Costs of Automated Thinking' (23 July 2019) www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.

[1] RxList, 'Provigil', (*RxList*, 16 June 2020) www.rxlist.com/provigil-drug.htm.

[2] 'How Aspirin Works', (1995) 15(1) *The University of Chicago Chronicle* http://chronicle.uchicago.edu/950817/aspirin.shtml.

discovery to a seemingly unlimited number of new areas of inquiry. Humanity's intellectual credit line is undergoing an extraordinary, unasked-for bump up in its limit.

To understand the problems with intellectual debt despite its boon, it helps first to consider a sibling: engineering's phenomenon of technical debt.

In the summer of 2012, the Royal Bank of Scotland applied a routine patch to the software it used to process transactions. It went poorly. Millions of customers could not withdraw their money, make payments, or check their balances.[3] One man was held in jail over a weekend because he couldn't make bail.[4] A couple was told to leave their newly-purchased home when their closing payment wasn't recorded.[5] A family reported that a hospital threatened to remove life support from their gravely ill daughter after a charity's transfer of thousands of dollars failed to materialize.[6] The problem persisted for days as the company tried to figure out what had gone wrong, reconstruct corrupted data, and replay transactions in the right order.

RBS had fallen victim to technical debt. Technical debt arises when systems are tweaked hastily, catering to an immediate need to save money or implement a new feature, while increasing long-term complexity. Anyone who has added a device every so often to a home entertainment system can attest to the way in which a series of seemingly sensible short-term improvements can produce an impenetrable rat's nest of cables. When something stops working, this technical debt often needs to be paid down as an aggravating lump sum – likely by tearing the components out and rewiring them in a more coherent manner.

Banks are particularly susceptible to technical debt because they computerized early and invested heavily in mainframe systems that were, and are, costly and risky to replace. Their core systems still process trillions of dollars using software written in COBOL, a programming language from the 1960s that's no longer taught in most universities.[7] Those systems are now so intertwined with Web extensions, iPhone apps, and systems from other banks, that figuring out how they work all over again, much less eliminating them, is daunting. Consulting firms like Accenture have charged firms like the Commonwealth Bank of Australia hundreds of millions to dollars to make a clean break.[8]

Two crashes of Boeing's new 737 Max 8 jets resulted in the worldwide grounding of its Max fleet. Analysis so far points to a problem of technical debt: The company raced to offer a more efficient jet by substituting in more powerful engines, while avoiding a comprehensive redesign in order to fit the Max into the original 737 genus.[9] That helped speed up production in a number of ways, including bypassing costly recertifications. But the new engines had a tendency to push the aircraft's nose up, possibly causing it to stall. The quick patch was to alter the aircraft's software to automatically push the nose down if it were too far up. Pilots were then expected to know what to do if the software itself acted wrongly for any reason, such as receiving

[3] M Hickman, 'NatWest and RBS Customers May Receive Compensation as 'Computer Glitch' Drags into Sixth Day' *Independent* (26 June 2012) www.telegraph.co.uk/finance/personalfinance/bank-accounts/9352573/NatWest-customers-still-unable-to-see-bank-balances-on-sixth-day-of-glitch.html.

[4] 'RBS Computer Problems Kept Man in Prison' (*BBC News*, 26 June 2012) www.bbc.com/news/uk-18589280.

[5] L Bachelor, 'NatWest Problems Stop Non-Customers Moving into New Home' *The Guardian* (22 June 2012) www.theguardian.com/money/2012/jun/22/natwest-problems-stop-non-customers-home?newsfeed=true.

[6] J Hall, 'NatWest Computer Glitch: Payment to Keep Cancer Girl on Life Support Blocked' *The Telegraph* (25 June 2012) www.telegraph.co.uk/finance/personalfinance/bank-accounts/9352532/NatWest-computer-glitch-payment-to-keep-cancer-girl-on-life-support-blocked.html.

[7] A Irrera, 'Banks Scramble to Fix Old Systems as IT 'Cowboys' Ride into Sunset' *Reuters* (11 April 2017) www.reuters.com/article/us-usa-banks-cobol/banks-scramble-to-fix-old-systems-as-it-cowboys-ride-into-sunset-idUSKBN17C0D8.

[8] Ibid.

[9] N Rivero 'A String of Missteps May Have Made the Boeing 737 Max Crash-Prone' (Quartz, 18 March 2019) https://qz.com/1575509/what-went-wrong-with-the-boeing-737-max-8/.

the wrong information about nose position from the plane's sensors. A small change occasioned another small change which in turn forced another awkward change, pushing an existing system into unpredictable behavior. While the needed overall redesign would have been costly and time consuming, and would have had its own kinks to work out, here the alternative of piling on debt contributed to catastrophe.

Enter a renaissance in long-sleepy areas of AI based on machine learning techniques. Like the complex systems of banks and aircraft makers, these techniques bear a quiet, compounding price that may not seem concerning at first, but will trouble us later. Machine learning has made remarkable strides thanks to theoretical breakthroughs, zippy new hardware, and unprecedented data availability. The distinct promise of machine learning lies in suggesting answers to fuzzy, open-ended questions by identifying patterns and making predictions. It can do this through, say, 'supervised learning', by training on a bunch of data associated with already-categorized conclusions. Provide enough labeled pictures of cats and non-cats, and an AI can soon distinguish cats from everything else. Provide enough telemetry about weather conditions over time, along with what notable weather events transpired, and an AI might predict tornadoes and blizzards. And with enough medical data and information about health outcomes, an AI can predict, better than the best physicians can, whether someone newly entering a doctor's office with pulmonary hypertension will live to see another year.[10]

Researchers have pointed out thorny problems of technical debt afflicting AI systems that make it seem comparatively easy to find a retiree to decipher a bank system's COBOL.[11] They describe how machine learning models become embedded in larger ones and can then be forgotten, even as their original training data goes stale and their accuracy declines.

But machine learning doesn't merely implicate technical debt. There are some promising approaches to building machine learning systems that, in fact, can offer some explanations[12] – sometimes at the cost of accuracy – but they are the rare exceptions. Otherwise, machine learning is fundamentally patterned like drug discovery, and it thus incurs intellectual debt. It stands to produce answers that work, without offering any underlying theory. While machine learning systems can surpass humans at pattern recognition and predictions, they generally cannot explain their answers in human-comprehensible terms. They are statistical correlation engines – they traffic in byzantine patterns with predictive utility, not neat articulations of relationships between cause and effect. Marrying power and inscrutability, they embody *Arthur C. Clarke's* observation that any sufficiently advanced technology is indistinguishable from magic.[13]

But here there is no *David Copperfield* or *Ricky Jay* who knows the secret behind the trick. No one does. Machine learning at its best gives us answers as succinct and impenetrable as those of a Magic 8-Ball – except they appear to be consistently right. When we accept those answers without independently trying to ascertain the theories that might animate them, we accrue intellectual debt.

[10] TJW Dawes and others, 'Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study' (2017) 283(2) *Radiology* https://pubmed.ncbi.nlm.nih.gov/28092203/.

[11] D Sculley and others, 'Hidden Technical Debt in Machine Learning Systems' (2018) 2 *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems* https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

[12] C Rudin, 'New Algorithms for Interpretable Machine Learning' (2014) KDD'14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining www.bu.edu/hic/2018/12/04/air-rudin/.

[13] AC Clarke, 'Hazards of Prophecy: The Failure of Imagination' in AC Clarke, *Profiles of the Future: An Enquiry into the Limits of the Possible* (1962).

Why is unpaid intellectual debt worrisome? There are at least three reasons, in increasing difficulty. First, when we don't know how something works, it becomes hard to predict how well it will adjust to unusual situations. To be sure, if a system can be trained on a broad enough range of situations, nothing need ever be unusual to it. But malefactors can confront even these supposedly robust systems with specially-crafted inputs so rare that they'd never be encountered in the normal course of events. Those inputs – commonly referred to as 'adversarial examples' – can look normal to the human eye, while utterly confusing a trained AI.

For example, computers used to be very bad at recognizing what was in photos. That made categorization of billions of online images for a search engine like Google Images inaccurate. Fifteen years ago the brilliant computer scientist *Luis von Ahn* solved the problem by finding a way for people, instead of computers, to sort the photos for free. He did this by making the 'ESP game'.[14] People were offered an online game in which they were shown images and asked to guess what other people might say was in them. When they were right, they earned points. They couldn't cash the points in for anything, but thousands of people played the game anyway. And when they did, their successful guesses became the basis for labeling images. Google bought *Luis's* game, and the field of human computation – employing human minds as computing power – took off.[15]

Today, Google's 'Inception' architecture – a specially-configured 'neural network' machine learning system – has become so good at image recognition that *Luis's* game is no longer needed to get people to label photos. We know how Inception was built.[16] But even its builders don't know how it gets a given image right. Inception produces answers, but not the kinds of explanations that the players of *Luis*'s game could offer if they were asked. Inception correctly identifies, say, cats. But it can't provide an explanation for what distinguishes a picture of a cat from anything else. And in the absence of a theory of cathood, it turns out that Inception can be tricked by images that any human would still immediately see as one of a cat.

MIT undergraduates were able to digitally alter the pixels of a standard cat photo to leave it visibly unchanged – and yet fool Google's state-of-the-art image detection engine into determining with 'hundred percent confidence' that it was looking at a picture of guacamole.[17] They then went a step further and painted a 3D-printed turtle in a way that looks entirely turtle-like to a human – and causes Google to classify it at every angle as a rifle.[18]

A system that had a discernible theory of whiskers and ears for cats, or muzzles for rifles, would be harder to fool – or at least would only be foolable along the lines that humans could be. But systems without theory have any number of unknown gaps in their accuracy. This is not just a quirk of Google's state-of-the-art image recognizer. In the realm of healthcare, systems trained to classify skin lesions as benign or malignant can be similarly tricked into flipping their previously-accurate judgments with an arbitrary amount of misplaced confidence,[19] and the prospect of

[14] L Von Ahn and L Dabbish, 'Labeling Images with a Computer Game' (2004) CHI'04 Proceedings of the 2004 Conference on Human Factors in Computing Systems 319.

[15] See J Zitthrain, 'Ubiquitous Human Computing' (2008) Oxford Legal Studies Research Paper No. 32 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1140445.

[16] C Szegedy and others 'Rethinking the Inception Architecture for Computer Vision' (2016) 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2818.

[17] A Ilyas and others, 'Black-box Adversarial Attacks with Limited Queries and Information' (*labsix*, 23 April 2018) www.labsix.org/limited-information-adversarial-examples/.

[18] A Athalye and others, 'Fooling Neural Networks in the Physical World with 3D Adversarial Objects' (*labsix*, 31 October 2017) www.labsix.org/physical-objects-that-fool-neural-nets/.

[19] SG Finlayson and others, 'Adversarial Attacks against Medical Deep Learning Systems' (2019) arXiv:1804.05296v3 https://arxiv.org/pdf/1804.05296.pdf.

triggering insurance reimbursements for such inaccurate findings could inspire the real world use of these techniques.[20]

The consistent accuracy of a machine learning system does not defend it against these kinds of attacks; rather, it may serve only to lull us into the chicken's sense that the kindly farmer comes every day with more feed – and will keep doing so. Charmed by its ready-to-hand predictive power, we will embed machine learning – like the asbestos of yesteryear – into larger systems, and forget about it. But it will remain susceptible to hijacking with no easy process for continuing to validate the answers it is producing, especially as we stand down the work of the human judges it will ultimately replace. Intellectual debt entails a trade-off for vulnerability that is easy to drift into just the way that technical debt does.

There is a second reason to worry as AI's intellectual debt piles up: the coming pervasiveness of machine learning models. Taken in isolation, oracular answers can generate consistently helpful results. But these systems won't stay in isolation. As AI systems gather and ingest the world's data, they'll produce data of their own – much of which will be taken up by still other AI systems. The New York Subway system has its own old-fashioned technical debt, as trains run through tunnels and switches whose original installers and maintainers have long moved on. How much more complicated would it be if that system's activities became synchronized with the train departures at Grand Central Terminal, and then new 'smart city' traffic lights throughout the five boroughs?

Even simple interactions can lead to trouble. In 2011, biologist *Michael Eisen* found from one of his students that an unremarkable used book – *The Making of a Fly: The Genetics of Animal Design* – was being offered for sale on Amazon by the lowest-priced seller for just over $1.7 million, plus $3.99 shipping.[21] The next cheapest copy weighed in at $2.1 million. The respective sellers were well established; each had thousands of positive reviews. When *Eisen* visited the page the next day, the prices had gone up yet further. As each day brought new increases from the sellers, *Eisen* performed some simple math: Seller A's price was consistently 99.83% that of Seller B. And Seller B's price was, each day, adjusted to be 127.059% that of Seller A.

Eisen figured that Seller A had a copy of the book and, true to the principles of Economics 101, was seeking to offer the lowest price of all sellers by slightly undercutting the next cheapest price. He then surmised that Seller B did not have a copy of the book, so priced it higher – and was then waiting to see if anyone bought the more expensive copy anyway. If so, Seller B could always get it from Seller A and direct delivery of the book to the lazy buyer, pocketing a handsome profit without having to actually personally package and mail anything.

Each seller's strategy is rational – and while algorithmic, surely involved no sophisticated machine learning at all. Even those straightforward strategies collided to produce manifestly irrational results. The interaction of thousands of machine learning systems in the wild promises to be much more unpredictable.

The financial markets provide an obvious breeding ground for this type of problem – and one in which cutting-edge machine learning is already being deployed today. In 2010, a 'flash crash' driven by algorithmic trading wiped more than $1 trillion from the major US indices – for thirty-six minutes. Last fall, JPMorgan analyst *Marko Kolanovic* shared a short analysis within a 168-page market report that suggested it could readily happen again, as more investing becomes

[20] SG Finlayson and others, 'Adversarial Attacks on Medical Machine Learning' 363 *Science* 1287.

[21] M Eisen, 'Amazon's $23,698,655.93 Book about Flies' it is NOT junk (22 April 2011) www.michaeleisen.org/blog/?p=358.

passive rather than active, and simply tied to indices.[22] Unlike technical debt, whose borrowing is typically attributable to a particular entity that is stewarding a system, intellectual debt can accumulate in the interstices where systems bump into each other without formally interconnecting.

A third, and most profound, issue with intellectual debt is the prospect that it represents a larger movement from basic science towards applied technology, one that threatens to either transform academia's investigative rigors or bypass them entirely.[23] Unlike, say, particle accelerators, the tools of machine learning are as readily taken up by private industry as by universities. Indeed, the kind and volume of data that will produce useful predictions is more likely to be in Google and Facebook's possession than at the MIT computer science department or Media Lab. Industry may be perfectly satisfied with answers that lack theory. But when those answers aren't themselves well publicized, much less the AI tools that produce them, intellectual debt will build in societal pockets far away from the academics who would be most interested in backfilling the theory. And an obsession only with answers – represented by a shift in public funding[24] of research to orient around them – can in turn steer even pure academics away from paying off the intellectual debt they might find in the world, and instead towards borrowing more.

One researcher in the abstruse but significant field of 'protein folding' recently wrote an essay exploring his ambivalence about what it means to be a scientist after a machine learning model was able to, well, fold proteins in ways that only humans had previously been able to achieve.[25] He told one publication: 'We've had this tendency as a field to be very obsessed with data collection. The papers that end up in the most prestigious venues tend to be the ones that collect very large data sets. There's far less prestige associated with conceptual papers or papers that provide some new analytical insight.'[26]

It would be the consummate pedant who refused to take a life-saving drug simply because no one knew how it worked. At any given moment an intellectual loan can genuinely be worth taking out. But as more and more drugs with unknown mechanisms of action proliferate – none of them found in the wild – the number of tests to uncover untoward interactions must scale exponentially. In practice, these interactions are simply found once new drugs find their way to the market and bad things start happening, which partially accounts for the continuing cycle of introduction-and-abandonment of drugs. The proliferation of machine learning models and their fruits makes that problem escape the boundaries of one field.

So, what should we do? First, we need to know our exposure. As machine learning and its detached answers rightfully blossom, we should invest in a collective intellectual debt balance sheet. Debt is not only often tolerable, but often valuable – it leverages what we can do. Just as a little technical debt in a software system can help adapt it to new uses without having to

[22] T Heath, 'The Warning from JPMorgan about Flash Crashes Ahead' *The Washington Post* (5 September 2018) www.washingtonpost.com/business/economy/the-warning-from-jpmorgan-about-flash-crashes-ahead/2018/09/05/25b1f90a-b148-11e8-a20b-5f4f84429666_story.html.

[23] K Birchard and J Lewington 'Dispute Over the Future of Basic Research in Canada' *The New York Times* (16 February 2014) www.nytimes.com/2014/02/17/world/americas/dispute-over-the-future-of-basic-research-in-canada.html.

[24] T Caulfield 'Should Scientists Have to Always Show the Commercial Benefits of Their Research?' (*Policy Options*, 1 December 2012) https://policyoptions.irpp.org/magazines/talking-science/caulfield/.

[25] M AlQuraishi 'AlphaFold @ CASP13: "What Just Happened?"' *Some Thoughts on a Mysterious* Universe, (9 December 2018) https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/#comment-26005.

[26] S Samuel, 'How One Scientist Coped When AI Beat Him at His Life's Work' (Vox, 15 February 2019) www.vox.com/future-perfect/2019/2/15/18226493/deepmind-alphafold-artificial-intelligence-protein-folding.

continually rebuild it, a measure of considered intellectual debt can give us a *Promethean* knowledge boost, and then signpost a research agenda to discover the theory that could follow.

For that, we need the signposts. We must keep track of just where we've plugged in the answers of an alien system, rather than tossing crumpled IOUs into a file cabinet that could come due without our noticing. Not all debt is created equal. When the stakes are low, such as the use of machine learning to produce new pizza recipes,[27] it may make sense to shut up and enjoy the pizza, never fretting about the theory behind what makes peanut butter and banana toppings work so well together on a pie. But when the stakes are higher, such as the use of AI to make health predictions and recommendations, we walk on untested ice when we crib the answers to a test rather than ever learning the underlying material. That it is near-irresistible to use the answers makes pursuing an accompanying theory all the more important.

To achieve a balance sheet for intellectual debt, we must look at current practices around trade secrets and other intellectual property. Just as our patent system requires public disclosure of a novel technique in exchange for protection against its copying by others, or the city building department requires the public availability of renovation plans for private buildings, we should explore academic mirroring and escrow of otherwise-hidden data sets and algorithms that achieve a certain measure of public use. That gives us a hope for building a map of debt – and a rapid way to set a research agenda to pay off debt that appears to have become particularly precarious.

Most important, we should not deceive ourselves into thinking that answers alone are all that matters: Indeed, without theory, they may not be meaningful answers at all. As associational and predictive engines spread and inhale ever more data, the risk of spurious correlations itself skyrockets. Consider one brilliant amateur's running list of very tight associations found,[28] not because of any genuine association, but because with enough data, meaningless, evanescent patterns will emerge. The list includes almost perfect correlations between the divorce rate in Maine and the per capita consumption of margarine, and between US spending on science, space, and technology and suicides by hanging, strangulation, and suffocation. At just the time when statisticians and scientists are moving to de-mechanize the use of statistical correlations,[29] acknowledging that the production of correlations alone has led us astray, machine learning is experiencing that success of the former asbestos industry relies on the basis of exactly those kinds of correlations.

Traditional debt shifts control, from borrower to lender, and from future to past, as later decisions are constrained by earlier bargains. Answers without theory – intellectual debt – also will shift control in subtle ways. Networked AI is moving decisions previously left by necessity to, say, a vehicle's driver into the hands of those tasked with designing autonomous vehicles – hence the ongoing hand-wringing around ethical trolley problems.[30] Society, not the driver, can now directly decide whom a car that has lost its brakes should most put at risk, including its passengers. And the past can now decide for the future: Cars can be programmed well ahead of time with decisions to be actualized later.

[27] 'Episode 2: Human-AI Collaborated Pizza' *How to Generate (Almost) Anything* (30 August 2018) https://howtogeneratealmostanything.com/food/2018/08/30/episode2.html.

[28] T Vigen, *Spurious Correlations* (2015).

[29] RL Wasserstein, AL Schirm, and NA Lazar, 'Moving to a World Beyond "p < 0.05"' (2019) 73(S1) *The American Statistician* www.tandfonline.com/doi/pdf/10.1080/00031305.2019.1583913?needAccess=true.

[30] G Marcus 'Moral Machines' *The New Yorker* (24 November 2012) www.newyorker.com/news/news-desk/moral-machines.

A world of knowledge without understanding becomes, to those of us living in it, a world without discernible cause and effect, and thus a world where we might become dependent on our own digital concierges to tell us what to do and when. It's a world where home insurance rates could rise or fall by the hour or the minute as new risks are accurately predicted for a given neighborhood or home. The only way to make sense of that world might be to employ our own AIs to try to best position us for success with renter's insurance AIs ('today's a good day to stay home'); hiring AIs ('consider wearing blue'); or admissions AIs ('volunteer at an animal shelter instead of a homeless shelter'), each taking and processing inputs in inscrutable ways.

When we have a theory, we get advanced warning of trouble when the theory stops working well. We are called to come up with a new theory. Without the theory, we lose the autonomy that comes from knowing what we don't know.

Philosopher *David Weinberger* has raised the fascinating prospect that machine learning could help us tap into natural phenomena that themselves don't avail themselves of any theory to begin with.[31] It's possible that there are complex but – with enough computing power – predictable relationships in the universe that simply cannot be boiled down to an elegant formula like Newton's account of gravity taught in high schools around the world, or Einstein's famed insight about matter, energy, and the speed of light. But we are soon to beat nature to that complex punch: with AI, in the name of progress, we will build phenomena that can only be predicted, while never understood, by other AI.

That is, we will build models dependent on, and in turn creating, underlying logic so far beyond our grasp that they defy meaningful discussion and intervention. In a particularly fitting twist, the surgical procedure of electrical deep brain stimulation has advanced through trial-and-error – and is now considered for the augmentation of human thinking, 'cosmetic neurosurgery'.[32]

Much of the timely criticism of AI has rightly focused on ways in which it can go wrong: it can create or replicate bias; it can make mistakes; it can be put to evil ends. Alongside those worries belongs another one: what happens when AI gets it right, becoming an Oracle to which we cannot help but to return and to whom we therefore become bonded.

[31] D Weinberger, 'Optimization over Explanation' (*Berkman Klein Center*, 28 January 2018) https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d.

[32] N Lipsman and AM Lozano, 'Cosmetic Neurosurgery, Ethics, and Enhancement' (2015) 2 *The Lancet Psychiatry* 585.

PART III

Responsible AI Liability Schemes

12

Liability for Artificial Intelligence

The Need to Address Both Safety Risks and Fundamental Rights Risks

Christiane Wendehorst

I. INTRODUCTION

On 21 April 2021, the European Commission published its package of measures on a European approach to artificial intelligence (AI), consisting of a communication,[1] accompanied by an updated Coordinated Plan on AI[2] and a proposal for a horizontal regulation (Artificial Intelligence Act, AIA)[3] with nine annexes. This package is the first of three inter-related legal initiatives announced by the Commission with the aim of making Europe a safe and innovation friendly environment for the development of AI. This first initiative aims to establish a European legal framework for AI to address fundamental rights and safety risks specific to AI systems. The second initiative is the revision of sectoral and more horizontal safety legislation. A proposal for a new Machinery Regulation[4] with eleven annexes was already published on the same day as the AI package, addressing an important aspect of AI usually referred to as 'robotics', and a proposal for a new General Product Safety Regulation[5] followed soon after. Parliament and Council are currently preparing both files for the trilogues. Finally, the third initiative announced is the introduction of EU rules to address liability issues related to new technologies, including AI systems. The Public Consultation for this initiative has already been closed and a proposal is planned for the third quarter of 2022.[6] This third initiative will comprise measures adapting the liability framework to the challenges of new technologies, including AI, to ensure that victims

[1] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Fostering a European Approach to Artificial Intelligence' COM (2021) 205 final.

[2] European Commission, 'Annex to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. New Coordinated Plan on AI 2021 Review' COM (2021) 205 final.

[3] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' COM (2021) 206 final.

[4] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on Machinery Products' COM (2021) 202 final.

[5] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and Repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council' COM (2021) 346 final.

[6] European Commission, 'Civil Liability: Adapting Liability Rules to the Digital Age and Artificial Intelligence' https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Civil-liability-adapting-liability-rules-to-the-digital-age-and-artificial-intelligence_en; this chapter was written in spring 2021, only certain sections have been updated.

who suffer damage to their life, health, or property as a result of new technologies have access to the same compensation as victims of other technologies. In the Inception Impact Assessment, a revision of the Product Liability Directive (PLD),[7] and a legislative proposal with regard to the liability for certain AI systems are identified as policy options.[8]

Given that liability for AI and other emerging digital technologies had been on the agenda for some time, it may come as a surprise that liability legislation figures last on the agenda. An Expert Group on Liability and new Technologies was established in 2018. It was divided into two formations, one dealing specifically with the PLD and being largely dominated by stakeholders, the other – the so-called New Technologies Formation (EG-NTF) – having a broader mandate and consisting mainly of academics.[9] Only the NTF ever published an official written report,[10] which then served, *inter alia*, as a basis for the European Commission's report on the safety and liability implications of AI, the Internet of Things (IoT), and robotics[11] of 19 February 2020, which formed part of the 2020 AI package and accompanied the Commission White Paper on AI.[12]

A major driver of activities in the field of liability has certainly been the European Parliament. After its first resolution in 2017,[13] which included the much-quoted and much-criticised plea for electronic personhood,[14] the European Parliament passed another resolution on 20 October 2020 that includes a full-fledged 'Proposal for a Regulation of the European Parliament and of the Council on liability for the operation of AI systems'.[15] This proposal is certainly much more mature than the 2017 resolution and bears a striking resemblance to policy considerations made within parts of the European Commission.

Whether the Commission will follow the recommendations of Parliament or take a different approach remains yet to be seen. Because AI liability is a subject matter that might be addressed

[7] Council Directive 85/374/EEC of 25 July 1985 on the Approximation of the Laws, Regulations and Administrative Provisions of the Member States Concerning Liability for Defective Products [1985] OJ L 2010/29; see European Commission, 'Commission Staff Working Document. Evaluation of Council Directive 85/374/EEC of 25 July 1985' SWD (2018) 157 final.

[8] European Commission, 'Adapting Liability Rules to the Digital Age and Artificial Intelligence' Inception Impact Assessment (Ares(2021)4266516).

[9] European Commission, 'Register of Commission Expert Groups, Expert Group on Liability and New Technologies (E03592)' (*European Commission*, 9 March 2018) https://ec.europa.eu/transparency/expert-groups-register/screen/expert-groups/consult?do=groupDetail.groupDetail&groupID=3592&Lang=NL.

[10] Directorate-General for Justice and Consumers, 'Liability for Artificial Intelligence and Other Emerging Digital Technologies' (*European Commission*, 27 November 2019) https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en/format-PDF (hereafter 'NTF Expert Group').

[11] European Commission, 'Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics' COM (2020) 64 final.

[12] European Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' COM (2020) 65 final.

[13] European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, P8_TA (2017)0051 (hereafter EP Resolution on Civil Law Rules on Robotics).

[14] G Wagner, 'Robot Liability' in S Lohsse, R Schulze, and D Staudenmayer (eds), *Liability for Artificial Intelligence and the Internet of Things* (2019) 44 *et seq*; BA Koch, 'Product Liability 2.0: Mere Update or New Version?' in S Lohsse, R Schulze and D Staudenmayer (eds), *Liability for Artificial Intelligence and the Internet of Things* (2019) (hereafter Koch, 'Product Liability 2.0: Mere Update or New Version?'); G Spindler, 'Roboter, Atomation, künstliche Intelligenz, selbst-steuernde Kfz – Braucht das Recht neue Haftungskategorien?' (2015) CR 766, 773; H Eidenmüller, 'The Rise of Robots and the Law of Humans' (2017) ZEuP 765, 774 *et seq*; R Schaub, 'Interaktion von Mensch und Maschine' (2017) JZ, 342, 345.

[15] European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence (2020/2014(INL)) P9_TA(2020)0276 (hereafter EP Resolution on a Civil Liability Regime for AI).

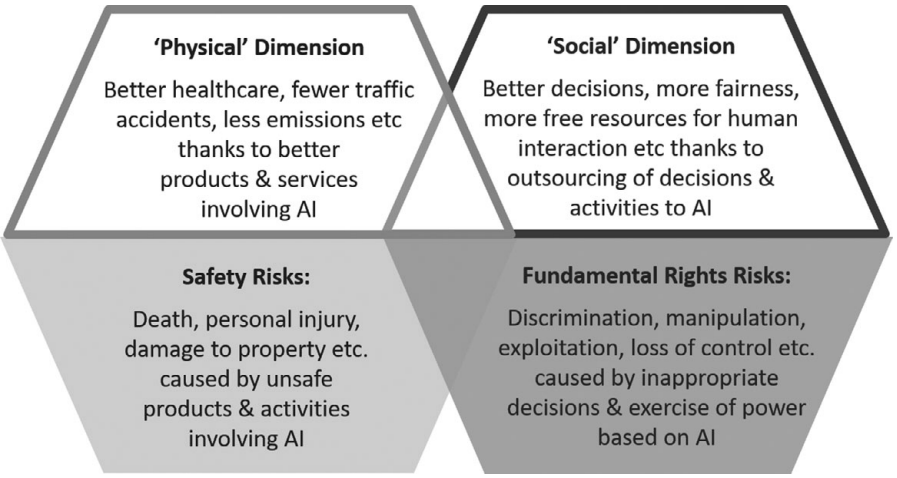

FIGURE 12.1 The 'physical' and the 'social' dimensions of risks associated with AI

within different regulatory and legal frameworks for which different Directorates General of the Commission and different Committees within the Parliament are responsible, the matter remains highly controversial. This paper analyses the different risks posed by AI, and why AI challenges existing liability regimes. It also explains the main solutions put forward so far and evaluates them, concluding that different solutions may be appropriate for different types of risk.

II. DIMENSIONS OF AI AND CORRESPONDING RISKS POSED

The challenges posed by AI and modern digital ecosystems in general – such as opacity ('black box-effect'), complexity, and partially 'autonomous' and unpredictable behaviour – are similar, irrespective of where and how AI is deployed. However, at a somewhat lower level of abstraction, the potential risks associated with AI usually appear to be falling into either of two dimensions: 'safety risks' and 'fundamental rights risks'.[16] These two types of risks are just the downside of our expectations of AI and of the promises made by those developing and deploying the technology, that is, that AI will both help by improving health and saving lives and the climate, and assist us in making better decisions, enhancing fairness, and developing into a better society (Figure 12.1).

1. *Traditional (Physical) Safety Risks*

Traditionally, death, personal injury, and damage to property have played a special role within safety and liability frameworks. These traditional types of risks can more specifically be described as 'physical' safety risks, but are normally referred to simply as 'safety risks'. These risks continue to play their very special role in the digital era, but the concept must be understood more broadly to include not only death, personal injury, and damage to property in the traditional sense, but

[16] In previous publications, I have referred to the two types as 'physical' and 'social' risks, see e.g. JP Schneider and C Wendehorst, 'Response to the Public Consultation on the White Paper: On Artificial Intelligence: A European Approach to Excellence and Trust, COM(2020) 65 final' (ELI 2020); C Wendehorst and Y Duller, *Safety and Liability Related Aspects to Software* (*European Commission*, 2021) (hereafter Wendehorst and Duller, 'Safety and Liability') 26 *et seq*; C Wendehorst, 'Strict Liability for AI and Other Emerging Technologies' (2020) JETL (hereafter Wendehorst, 'Strict Liability') 150, 161 *et seq*.

also damage to data and to the functioning of other digital systems. Where, for example, the malfunctioning of software causes the erasure of important customer data stored by the data holder in some cloud space, this should have the same legal effect as the destruction of a hard disk drive or of paper files with customer data (which is not to say that all data should automatically be treated in exactly the same way as tangible property in the tort liability context).[17] Likewise, where tax management software causes the victim's customer management software to collapse, this must be considered a safety risk, irrespective of whether the customer management software was run on the victim's hard disk drive or somewhere in the cloud within a SaaS scheme. While this is unfortunately still disputed under national tort law,[18] any attempt to draw a line between data stored on a physical medium owned by the victim and data stored otherwise seems to be completely outdated and fails to recognise the functional equivalence of different forms of storage.

2. *Fundamental Rights Risks*

'Fundamental rights risks' are associated with the social dimension of AI. They include discrimination, exploitation, manipulation, humiliation, oppression, and similar undesired effects that are – at least primarily – non-economic (non-material) in nature and that are not just the result of physical harm (as the latter would be dealt with under traditional regimes of compensation for pain and suffering, etc). Such risks have traditionally been dealt with primarily by special legal regimes, such as data protection law, anti-discrimination law or, more recently, law against hate speech on the Internet and similar legal regimes.[19] There is also a growing body of tort law that deals specifically with the infringement of personality rights.[20] Even though the concept of 'fundamental rights' is focused on individual rights, the term 'fundamental rights risks' should be understood more broadly as encompassing also risks of a more collective nature, for example, risks for the rule of law, democracy, and freedom of expression in general.[21]

While the fundamental rights aspect and, therefore, the non-economic aspect of such risks is in the foreground, these risks can, of course, entail economic risks for the affected individual or for society as a whole. For instance, AI systems used for recruitment that favour male applicants create a social risk for female applicants by discriminating against them, but this also leads to adverse economic effects for the affected women.

[17] C Wendehorst, "Liability for Pure Data Loss" in E Karner and others (eds) *Festschrift für Helmut Koziol* (2020) 225 (hereafter Wendehorst, 'Liability for Pure Data Loss').

[18] See Wendehorst, 'Liability for Pure Data Loss' (n 17) 225; G Wagner, '§ 823' in FJ Säcker and others (eds), *Münchener Kommentar zum BGB* (8th ed. 2020) para 245 *et seq*; L Specht, *Konsequenzen der Ökonomisierung informationeller Selbstbestimmung* (2012) 230; F Faust, 'Digitale Wirtschaft: Analoges Recht: Braucht das BGB ein Update?' in Ständige Deputation des Deutschen Juristentages (ed), *Verhandlungen des 71. Deutschen Juristentages – Band I – Gutachten Teil A* (2016), 48.

[19] Regulation (EU) 2016/679, Article 82(1); Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L 373/37, Article 8(2); German Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG, BGBl I S 3352); French Anti-Hate Speech Law (Loi Avia 2020/766); Austrian Anti-Hate Speech Law (Hass-im-Netz-Bekämpfungs-Gesetz, HiNBG, BGBl I 2020/148); Proposal for a Regulation of the European Parliament and the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM (2020) 825 final.

[20] For an overview see G Brüggemeier, AC Ciacchi, and P O'Callaghan, *Personality Rights in European Tort Law* (2010).

[21] C Wendehorst, 'The Proposal for an Artificial Intelligence Act COM(2021) 206 from a Consumer Policy Perspective' (*Federal Ministry Republic of Austria for Social Affairs, Health, Care and Consumer Protection*, 2021) (hereafter Wendehorst, 'The Proposal for an AIA from a Consumer Policy Perspective'), 110.

3. Overlaps and In-Between Categories

The division between safety and fundamental rights risks is generally not always clear-cut and should not be overestimated. There are not only clear overlaps, but also a considerable grey area of a number of important risks. For instance, adverse psychological effects can be a very traditional safety risk,[22] where the effect is a diagnosed illness according to WHO criteria (such as depression), but also a fundamental rights risk that is associated with the social dimension of AI where the effect is not a diagnosed illness, but, for example, just stress or anxiety. It is not always easy to draw a line between the two.[23]

a. Cybersecurity and Similar New Safety Risks

Digitalisation has given rise to a number of very special risks that are not easy to classify. They are essentially safety risks, albeit safety risks of a nature that is somewhat in a grey zone between 'physical' and 'intangible'. Such special safety risks include the 'data security' aspect of data protection and privacy (i.e. prevention of data leaks), cybersecurity and harm to the network, and fraud or illegal collusion, to name but a few. They are recognised as relevant safety risks under selected pieces of safety legislation, in particular the Radio Equipment Directive (RED)[24] and the Medical Device Regulation (MDR).[25] Digital risks are also recognised in the Proposal for a Regulation on Machinery Products[26] and the Proposal for a Regulation on General Product Safety,[27] which are intended to replace the Directives currently in force. However, these (digital) risks will often primarily relate to the 'physical' dimension of safety, because data theft and manipulation or the breakdown of networks and other essential infrastructures will indirectly, at least in most cases, lead to damage to property in the broader sense or even threaten the health and life of persons.

b. Pure Economic Risks

Pure economic risks[28] are economic risks that are not just the result of the realisation of physical risks, such as personal injury or property damage. Where medical AI causes a surgery to fail, resulting in personal injury and consequently in hospitalisation, the costs of hospitalisation is an economic harm, but not a 'pure' economic harm because it results from the personal injury. Where, however, AI manipulates consumers and makes them buy overpriced products, the financial loss caused is not in any way connected with a safety risk and, therefore, qualifies as a pure economic risk (also referred to as immaterial harm). For pure economic risks to be

[22] Article 10:202(1) of the Principles of European Tort Law (hereafter PETL) prepared by the European Group on Tort Law http://egtl.org/PETLEnglish.html.

[23] C van Dam, *European Tort Law* (2006) (hereafter Van Dam, *European Tort Law*) 147.

[24] Article 3(3) Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the Harmonisation of the Laws of the Member States Relating to the Making Available on the Market of Radio Equipment and Repealing Directive 1999/5/EC [2014] OJ L 153/62.

[25] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, [2017] OJ L 117/1, Annex I, 14.2.

[26] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on Machinery Products' COM (2021) 202 final, Annex III, 1.1.9. and 1.2.1.

[27] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council' COM (2021) 346 final, Article 7(1)(h).

[28] PETL, Article 2:102(4); Van Dam, *European Tort Law* (n 20) 169.

considered legally relevant outside the realm of contractual liability, most legal systems require additional elements, such as fraud or other illegal behaviour or conduct that is considered socially inacceptable.[29] Pure economic risks, at least when legally relevant, might, therefore, be closer to fundamental rights risks.

III. AI AS A CHALLENGE TO EXISTING LIABILITY REGIMES

1. *Classification of Liability Regimes*

While extra-contractual liability law has – beyond product liability law and some few specific areas – so far largely been a matter for the Member States, and while there exists a broad variety of different liability regimes at national level, it is still possible to group liability regimes according to their general characteristics.

a. Fault Liability

Fault liability has been the most important pillar of extra-contractual liability in a majority of European jurisdictions.[30] Liability always requires a sufficient justification for shifting loss from the person who originally suffered the damage (the victim) to a person who caused the damage (the tortfeasor). In the case of fault liability, the fault of the tortfeasor, which is usually either intent or negligence with many different shades and gradations, such as gross negligence or recklessness, is the justification. If damage is caused by mere negligence, further conditions must usually be met, otherwise liability could potentially escalate indefinitely. Jurisdictions use different tools in order to keep liability within reasonable boundaries. Often, there is a requirement that the potential tortfeasor's conduct was somehow objectionable, that is, that it was either violating the law, or public policy, or infringing rights and legally protected interests whose absolute integrity is so vital that any kind of infringement must, per se, be considered as presumably unlawful. The latter is usually the case where human life, health, or bodily integrity are at stake or where the infringement concerns clearly defined property rights.[31]

b. Non-Compliance Liability

Liability may also be triggered by the infringement of particular laws or particular standards whose purpose includes the prevention of harm of the type at hand. We find this type of liability regime both at EU level and at national level. An example for non-compliance liability at EU level is Article 82 of the General Data Protection Regulation (GDPR),[32] which attaches liability to any infringement of the requirements set out by the GDPR. Further, yet very different, examples can be found in EU non-discrimination legislation such as Council Directive 2004/113/EC.[33] Non-discrimination law obliges Member States to introduce into their national legal systems the legal measures necessary to ensure real and effective compensation for loss and

[29] G Brüggemeier, *Tort Law in the European Union* (2nd ed. 2018) para 385; B Wininger and others (eds), *Digest of European Tort Law Volume 2: Essential Cases on Damage* (2011) 383 *et seq*.

[30] For a comparative report, see P Widmer (ed), *Unification of Tort Law: Fault* (2005).

[31] PETL, Article 2:102.

[32] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

[33] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services.

damage sustained by a person injured as a result of discrimination, in a way which is dissuasive and proportionate to the damage suffered. In this context, Member States must ensure that, when a plaintiff establishes facts from which it may be presumed that there has been direct or indirect discrimination, it shall be for the respondent to prove that there has been no breach of anti-discrimination law.[34] Another example of non-compliance liability can be found in the financial sector. Where issuers of a financial instrument do not publicly disclose inside information concerning them, they become liable for any damage caused by the failure to do so.[35]

At the national level, there may be both general clauses attaching liability to the infringement of protective statutory provisions[36] and specific liability regimes attaching liability to non-compliance with very particular standards. Non-compliance liability is always of an accessory nature, in other words, there needs to be a basic regime setting out in some detail the duties and obligations to be met in order to be considered compliant. It should also be noted that, under a number of national jurisdictions, efforts are being made to impose non-compliance liability only in cases where the potential tortfeasor was at fault.[37]

c. Defect and Mal-Performance Liability

A number of different liability regimes in jurisdictions in Europe may be described as types of 'defect liability' (or, in the case of services, 'mal-performance liability'), although this is certainly not a common technical term. In the extra-contractual realm, the most important form of defect liability is product liability, which has been harmonised by the Product Liability Directive (PLD).[38] Product liability does not require fault on the part of the producer, but it still requires a particular shortcoming in the producer's sphere, in that it requires that the product put into circulation was defective at the time when it left that sphere. The development risk defence (i.e. the defence relying on the fact that the defect, according to the state of the art in science and technology, could not have been detected when the product was put into circulation), which Member States were free to implement or not, moves product liability somewhat into the vicinity of fault liability.[39]

Product liability is only the most conspicuous form of defect liability and the one where the term 'defect' is in fact used. However, when looking more closely at liability regimes in national jurisdictions, it becomes apparent that there is a panoply of different forms of liability that are all based on the unsafe or otherwise objectionable state of a particular object within the liable person's sphere of control. Many of these forms of liability are somewhat at the borderline between fault liability and defect liability, as they are based on a presumption of fault, which the liable person is free to rebut under particular circumstances. Even some forms of vicarious liability under national law may be qualified, at a closer look, as forms of defect or mal-performance liability. For example, vicarious liability may be based on the generally 'unfit' nature of the

[34] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation [2005] OJ L 204/23, Article 18; Council Directive 2004/113/EC, Article 9; Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L 303/16, Article 10.

[35] Explicitly in sections 97 and 98 of the German Securities Trading Act.

[36] See, for example, section 823(2) of the German Civil Code (Bürgerliches Gesetzbuch, BGB) and section 1311 of the Austrian Civil Code (Allgemeines Bürgerliches Gesetzbuch, ABGB).

[37] J Fedtke and U Magnus in BA Koch and H Koziol (eds), Unification of Tort Law: Strict Liability (2002) 147.

[38] Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L 210/29; see for the implementation of the Directive in the Member States WH van Boom and others, 'Product Liability in Europe' in H Koziol and others (eds), *Product Liability Fundamental Questions in a Comparative Perspective* (2017) 255 *et seq*.

[39] NTF Expert Group (n 10) 27 *et seq*.

relevant auxiliary in terms of personality or skills,[40] or on the fact that the human auxiliary failed to meet a particular objective standard of care.

d. Strict Liability

The term 'strict liability', although often used with a broader meaning, should be reserved for such forms of liability that do not require any kind of defect or mal-performance but are more or less based exclusively on causation. At a closer look, some further requirements beyond causation may have to be met, such as that the risk that ultimately materialised was within the range of risks covered by the relevant liability regime, and there may possibly be defences, such as a *force majeure* defence.[41]

Strict liability is usually imposed only in situations where significant and/or frequent harm may occur despite the absence of any fault or any identifiable defect, mal-performance, or other non-compliance. It is also imposed where such elements would be so difficult for the victim to prove that requiring such proof would lead to massive under-compensation or inefficiency. Paradigm cases are the operation of aircraft, railways, ships, or motor vehicles, although solutions in the EU Member States differ, as does the attitude towards a 'general clause' of strict liability for unforeseen but parallel cases.[42] While there are also examples in national law where something close to strict liability is extended to all objects,[43] this is more or less exceptional and often narrowed down by case law.

2. *Challenges Posed by AI*

The mass rollout of AI and related technologies poses numerous challenges to existing liability regimes. Some of these challenges have their origin in interconnectedness, which is not strictly related to AI, but to digital ecosystems more generally. Other challenges are truly specific to AI.

a. Liability for the Materialisation of Safety Risks

(1) 'COMPLEXITY', 'OPENNESS', AND 'VULNERABILITY' OF DIGITAL ECOSYSTEMS With enhanced connectivity and data flows in the Internet of Things (IoT), everything potentially affects the behaviour of everything, and it may become close to impossible for a victim to prove what exactly caused the damage ('complexity'[44]). For example, where a smart watering system for the garden floods the premises, this may be the effect of the watering system itself being unsafe, but there might also have been an issue with a humidity sensor bought separately, or with the weather data supplied by another provider.

'Openness'[45] means the fact that components are not static but dynamic and are subject to frequent or even continuous change. Products change their safety-relevant features after the product has been put into circulation, for example through the online provision of updates as well as through a variety of different data feeds and cloud-based digital services. This, in fact,

[40] See e.g. section 1315 of the Austrian Civil Code (ABGB).
[41] PETL, Article 7:102(1 a) and Article 5:101 (1); BA Koch and H Koziol, 'Country Report Austria' in BA Koch and H Koziol (eds), *Unification of Tort Law: Strict Liability* (2002) 12, 15, 19.
[42] BA Koch and H Koziol, 'Comparative Conclusions' in BA Koch and H Koziol (eds), *Unification of Tort Law: Strict Liability* (2002) 395 *et seq*.
[43] Responsabilité du fait des choses, Article 1242 Code civil.
[44] NTF Expert Group (n 10) Key Finding no 1(a) 32 *et seq*.
[45] NTF Expert Group (n 10) Key Finding no 1(c) 32 *et seq*.

means that a victim may not get compensation under liability regimes such as the PLD which exclusively refer to the point in time when a product was first put into circulation.[46]

Connectivity also gives rise to increased 'vulnerability',[47] due to cyber security risks and privacy risks as well as a number of related risks, such as risks of fraud. However, as has been demonstrated by the short survey of existing liability regimes, such risks are not necessarily covered by liability because of a general focus on risks of a 'physical' nature such as death, personal injury, or property damage.

(II) 'AUTONOMY' AND 'OPACITY' AI adds further challenges to an already challenging picture through the features of 'autonomy' and 'opacity'. The term 'autonomy', whose use with regard to machines has often been criticised because of its inextricable link with the free human will, refers to a certain lack of predictability as far as the reaction of the software to unseen instances is concerned. It is in particular when coding of the software has occurred wholly or partially with the help of machine learning[48] that it is difficult to predict how the software will react to each and every situation in the future.[49]

While unpredicted behaviour in new situations nobody had ever thought about may also occur with software of a traditional kind, algorithms created with the help of machine learning cannot easily be analysed, especially not when sophisticated methods of deep learning have been used. This 'opacity' of the code[50] ('black box effect') means that it is not easy to explain why an AI behaved in a particular manner in a given situation, and even less easy to trace that behaviour back to any feature which could be called a 'defect' of the code or to any shortcoming in the development process.

Both autonomy and opacity make it difficult to trace harm back to any kind of intent or negligence on the part of a human actor, which is why fault liability is not an ideal response to risks posed by AI. However, it is also clear that emerging digital technologies, notably AI, make it increasingly difficult to identify a defect due to the autonomy of software and software-driven devices as well as the opacity of the code, which means that defect liability may not be a wholly satisfactory response either.

(III) STRICT AND VICARIOUS LIABILITY AS POSSIBLE RESPONSES As the 'autonomy' and 'opacity' of AI may give rise to exactly the kind of difficulties strict liability is designed to overcome,[51] the further extension of strict liability to AI applications is increasingly being discussed. This would, at the same time, solve some of the problems associated with 'complexity', 'openness', and 'vulnerability' that come with the IoT. For instance, where it is unclear whether the flooding of the premises was due to a defect of the watering system itself, a humidity sensor, or a data feed, it is still clear that the water itself came from the pipes. Thus, if the

[46] Council Directive 85/374/EEC, Article 6(1)(c), Article 7(b); P Machnikowski, 'Conclusions' in P Machnikowski (ed), *European Product Liability: An Analysis of the State of the Art in the Era of New Technologies* (2016) 669, 695.

[47] NTF Expert Group (n 10) Key Finding no 1(g) 32 *et seq*.

[48] Article 3(a) of the EP Resolution on a Civil Liability Regime for AI (n 15) defines 'AI-system' as 'a system that is either software-based or embedded in hardware devices, and that displays behaviour simulating intelligence by, inter alia, collecting and processing data, analysing and interpreting its environment, and by taking action, with some degree of autonomy, to achieve specific goals'.

[49] NTF Expert Group (n 10) Key Finding nos 1(d) and (e) at 32, 33.

[50] NTF Expert Group (n 10) Key Finding no 1(b) 32, 33.

[51] See also NTF Expert Group (n 10) Key Finding no 9, 39 *et seq*.

legislator introduced strict liability for smart watering systems, this would mean that whoever is the addressee of this strict liability (e.g. the operator or the producer of the watering system) would have to compensate victims for harm suffered from water spread by the system. There have been extensive discussions as to who is the right addressee of liability, and as to which types of risks should ultimately be covered.[52]

Similar effects may be achieved by extending vicarious liability to situations where sophisticated machines are used in lieu of human auxiliaries. Otherwise, parties could escape liability by outsourcing a particular task to a machine rather than to a human auxiliary.[53]

For some time, there has been a debate whether to recognise that highly sophisticated robots, and software agents may themselves be the addressees of liability. The idea of 'electronic personhood' was fuelled by a 2017 European Parliament resolution,[54] but the proposal was met with a great deal of resistance since.[55] Some of the resistance had its roots in ethical considerations,[56] but there are also practical flaws. Being the addressee of liability, AI systems would have to be equipped with funds or with equivalent insurance, which means that electronic personhood is more an additional complication than a solution.[57] Another radical solution proposed is that of replacing liability schemes altogether by insurance or funds so that those suffering harm from AI would be compensated by a general compensation scheme to which, in particular, producers and maybe professional users would be contributing.[58] However, it is meanwhile broadly accepted that such schemes could realistically only be implemented for very particular applications and fields, such as connected driving, but not across the board for a general purpose technology such as AI.[59]

b. Liability for the Materialisation of Fundamental Rights Risks

The main challenge to existing liability schemes is the fact that they are entirely inadequate to address the challenges posed by AI, due to their focus on safety risks. Where fundamental rights risks posed by AI materialise, there is often no fault on the part of those deploying the AI, and it may be close to impossible for a victim to prove that there was fault on the part of the producer. Defect liability, at least as it currently exists under the PLD and under national legal regimes, is entirely focussed on traditional safety risks. This holds true to an even greater extent for strict liability, which, for the time being, is almost exclusively restricted to physical risks. Further, extending vicarious liability to situations where sophisticated machines are deployed in lieu of

[52] Wendehorst and Duller, 'Safety and Liability' (n 16) 93 *et seq*; Wendehorst, 'Strict Liability'(n 16) 165 *et seq*.

[53] NTF Expert Group (n 10) Key Findings nos 18 and 19, 45 *et seq*; H Zech, 'Entscheidungen digitaler autonomer Systeme: Empfehlen sich Regelungen zu Verantwortung und Haftung?' in Ständige Deputation des Deutschen Juristentages (ed), *Verhandlungen des 73. Deutschen Juristentages – Band I – Gutachten Teil A* (2020) (hereafter Zech, 'Entscheidungen digitaler autonomer Systeme') 76 *et seq*.

[54] EP Resolution on Civil Law Rules on Robotics (n 13).

[55] See e.g. the Open Letter to the European Commission Artificial Intelligence and Robotics (2018) www.robotics-openletter.eu/.

[56] Data Ethics Commission, Opinion of the German Data Ethics Commission (*BMJV*, 2019) 219 www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html.

[57] NTF Expert Group (n 10) Key Finding no 8, 36 *et seq*.

[58] EP Resolution on Civil Law Rules on Robotics (n 13), paras 57, 59; G Borges, 'New Liability Concepts: The Potential of Insurance and Compensation Funds' in S Lohsse, R Schulze, and D Staudenmayer (eds), *Liability for Artificial Intelligence and the Internet of Things* (2019) 148 *et seq*; Zech, 'Entscheidungen digitaler autonomer Systeme' (n 53) 105 *et seq*.

[59] J Hanisch, 'Zivilrechtliche Haftungskonzepte für Robotik' in E Hilgendorf (ed), *Robotik im Kontext von Recht und Moral* (2014) 43; J Eichelberger, 'Zivilrechtliche Haftung für KI und Smarte Robotik' in M Ebers and others (eds), *Künstliche Intelligenz und Robotik* (2020) 198.

human auxiliaries[60] may help also with regard to fundamental rights risks, as long as there is a basis for liability of the hypothetical human auxiliary. Non-compliance liability might possibly be an option, but beyond non-discrimination law, the GDPR, and unfair commercial practices law there is currently not much of a general compliance regime that could serve as a 'backbone' for AI liability. Of course, this 'backbone' could theoretically be created by the emerging AI safety legislation. This is why it is essential to analyse this legislation.

IV. THE EMERGING LANDSCAPE OF AI SAFETY LEGISLATION

While the debate on challenges posed by AI to existing liability regimes is still ongoing, the landscape of AI-relevant product safety law is already changing rapidly, as illustrated by the proposals for a new Machinery Regulation and for the AIA. It is important to understand the emerging safety regimes, because it is only against their background that liability regimes specifically tailored to AI can be properly designed.

1. *The Proposed Machinery Regulation*

a. General Aims and Objectives

The proposed Machinery Regulation aims at modernising the existing machinery safety regime harmonised by the Machinery Directive,[61] in particular with regard to new technologies. This concerns potential risks that originate from a direct human-robot collaboration, risks originating from connected machinery, the phenomenon that software updates affect the 'behaviour' of the machinery after its placing on the market, and the problems associated with risk assessment on machine learning applications before the product is placed on the market. Also, the current regime harmonised by the Machinery Directive still foresees a driver or an operator responsible for the movement of a machine, but fails to set up requirements for autonomous machines. Needless to say, there were also developments to consider and inconsistencies to fix that were not directly related to software and AI. The current list of high-risk machines in Annex I to the Directive was elaborated 15 years ago and is urgently in need of an update.

b. Qualification As High-Risk Machinery

Within the product safety framework for machinery, the qualification of machinery products as high-risk machinery plays an important role. Amongst others, in Annex I, all software ensuring safety functions, including AI systems, and all machinery embedding AI systems ensuring safety functions has been added to the list of high-risk machinery.[62] The fact that all safety components that are software components, and all machinery embedding AI for the purpose of ensuring safety functions, are now included in the list of high-risk machinery automatically means under the proposed Machinery Regulation that, for this kind of machinery, only third party certification will be accepted, even when manufacturers apply the relevant harmonised standards.

A machinery product is included in the list of high-risk machinery products if it poses a particular risk to human health. The notion of 'safety' therefore seems to refer exclusively to risks

[60] NTF Expert Group (n 10) Key Findings nos 18 and 19, 45 *et seq*.
[61] Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC [2006] OJ L 157/24.
[62] Annex I to COM (2021) 202 final, nos 24 and 25.

of a physical nature. The risk posed by a certain machinery product is, according to Article 5(3) of the Proposal, established based on the combination of the probability of occurrence of harm and the severity of that harm. Factors to be considered in determining the probability and severity of harm include the degree to which each affected person would be impacted by the harm, the number of persons potentially affected, the degree of reversibility of the harm, and indications of harm that have been caused in the past by machinery products which have been used for relevant purposes. However, there are also factors that go more in the direction of 'fundamental rights risks', such as the degree to which potentially affected parties are dependent on the outcome produced by the machinery product, and the degree to which potentially affected parties are in a vulnerable position vis-à-vis the user of the machinery product.

c. Essential Health and Safety Requirements

The essential health and safety requirements that must be met for conformity of high-risk machinery are listed in Annex III. Where machinery uses AI for safety functions, the conformity assessment must consider hazards that may be generated during the lifecycle of the machinery as an intended evolution of its fully or partially evolving behaviour or logic.[63] As far as human-machine collaboration is concerned, a machinery product with fully or partially evolving behaviour or logic that is designed to operate with varying levels of autonomy must be adapted to respond to people adequately and appropriately; this must occur verbally through words or nonverbally through gestures, facial expressions, or body movement. It must also communicate its planned actions (what it is going to do and why) to operators in a comprehensible manner.[64]

Largely, however, AI-specific aspects are referred to in the future AIA, that is, where the machinery product integrates an AI system, the machinery risk assessment must consider the risk assessment for that AI system that has been carried out pursuant to the AIA.[65]

2. *The Proposed Artificial Intelligence Act*

a. General Aims and Objectives

The AIA Proposal of 21 April 2021 aims at ensuring that AI systems placed on the Union market and used in the Union are safe and respect existing law on fundamental rights and Union values, and at enhancing governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems. At the same time, efforts are being made to ensure legal certainty in order to facilitate investment and innovation in AI and to facilitate the development of a single market for AI applications and prevent market fragmentation. The AIA is complementary to existing data protection law (in particular the GDPR and the Law Enforcement Directive[66]), non-discrimination law, and consumer protection law.

As regards high-risk AI systems, which are safety components of products, the AIA will be integrated into the existing and future product safety legislation. For high-risk AI systems related

[63] Annex III to COM (2021) 202 final, no 1(c).
[64] Annex III to COM (2021) 202 final, no 1.3.7.
[65] Annex III to COM (2021) 202 final, no 1(c).
[66] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L 119/89.

to products covered by the New Legislative Framework (NLF) legislation (e.g. machinery, medical devices, toys), the requirements for AI systems set out in the AIA will be checked as part of the existing conformity assessment procedures under the relevant NLF legislation.[67] The latter may, at the same time, include further AI-specific requirements relevant only in a particular sector. AI systems related to products covered by relevant 'old approach' legislation (e.g. aviation, motor vehicles)[68] are not directly covered by the AIA, though.[69]

b. The Risk-Based Approach

The AIA Proposal follows a risk-based approach, differentiating between uses of AI that create an unacceptable risk, a high risk, a limited risk, and a low or minimal risk.

(i) PROHIBITED AI PRACTICES Title II lists some narrowly defined AI systems whose use is considered unacceptable as contravening EU values and violating fundamental rights, such as manipulation through subliminal techniques or exploitation of group-specific vulnerabilities (e.g. children) in a manner that is likely to cause affected persons psychological or physical harm. The Proposal also prohibits general-purpose social scoring by public authorities and, subject to a range of exceptions, the use of 'real time' remote biometric identification systems in publicly accessible spaces for law enforcement purposes.[70]

(ii) HIGH-RISK AI SYSTEMS Title III contains mandatory essential requirements for AI systems qualified as 'high-risk' AI systems, defined as systems that create a high risk to the health and safety or fundamental rights of natural persons. There are two main categories of high-risk AI systems: AI systems used as a safety component of products that are subject to third party *ex ante* conformity assessment under NLF legislation listed in Annex II; and other stand-alone AI systems explicitly listed in Annex III. The systems listed in Annex III, as it currently stands, more or less exclusively address fundamental rights risks. This includes biometric identification and categorisation of natural persons; education and vocational training; employment; workers management and access to self-employment; access to, and enjoyment of, essential private services, public services, and benefits; law enforcement; migration, asylum and border control management; and administration of justice and democratic processes. The only exception is the 'management and operation of critical infrastructure'[71] as the latter poses a systemic risk of a more physical nature rather than a fundamental rights risk.

The Commission may, from time to time, expand the list of high-risk AI systems used within certain pre-defined areas, by applying a set of criteria and risk assessment methodology. The risk assessment criteria listed in Article 7(2) are similar to those listed in the relevant Article of the proposed Machinery Regulation,[72] with two main exceptions: Reference is not only made to risks for the health of persons, but also to risks for the 'health and safety or … fundamental rights'. Also, an additional criterion to consider is the extent to which existing Union legislation already provides for effective measures of redress in relation to the risks posed by an AI system (with the exclusion of claims for damages) and the existence of effective measures to prevent or

[67] Article 6(1)(b); Recital 63 COM (2021) 202 final.
[68] Annex II section B to COM (2021) 202 final.
[69] Article 2(2)(2) COM (2021) 202 final.
[70] Wendehorst, 'The Proposal for an AIA from a Consumer Policy Perspective' (n 21) 75.
[71] Annex III to COM (2021) 202 final, no 2.
[72] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on Machinery Products' COM (2021) 202 final, Article 5(3).

substantially minimise those risks. For the purpose of future classification of additional AI systems as 'high-risk' systems, safety risks and fundamental rights risks are treated in the same manner and are not dealt with separately.

(iii) AI SYSTEMS SUBJECT TO SPECIFIC TRANSPARENCY OBLIGATIONS Title IV is devoted to AI systems that are subject to enhanced transparency obligations. This concerns, for example, AI systems that may be mistaken for human actors, deep fakes, emotion recognition systems, and biometric categorisation systems.[73] It is important to note, though, that Titles III and IV are not mutually exclusive, i.e. an AI system that qualifies as a 'high-risk' system for the purpose of Title III may still fall under IV as well.

c. Legal Requirements and Conformity Assessment for High-Risk AI Systems

Legal requirements set out in Title III for high-risk AI systems address data and data governance, documentation and record keeping, transparency and provision of information to users, human oversight, robustness, accuracy, and security. By and large, and with regard to the AI system, the same requirements apply irrespective of whether what is at stake is the safety component of a toy robot or a connected household device falling under the RED, or an AI system intended to be used for the selection and evaluation of applicants in the course of a recruitment procedure. This may not be particularly convincing, because the safety requirements with regard to the toy robot or the connected household device are very different to the safety requirements with regard to the recruitment software. However, due to the general nature of the requirements and obligations listed in the Proposal, it may still be the better choice to deal with the two risk categories under identical provisions.

Obligations with regard to these requirements are largely placed on producers (called 'providers') of high-risk AI systems, but proportionate obligations are also placed on (professional) users and other participants across the AI value chain (such as importers, distributors, and authorised representatives) consistent with other modern product safety legislation. The Proposal sets out a framework for notified bodies to be involved as independent third parties in conformity assessment procedures. AI systems used as safety components of products regulated under the NLF, such as machinery or toys, are subject to the same compliance and enforcement mechanisms of the products of which they are a component, but in the course of applying these mechanisms the requirements imposed by the AIA must be ensured as well. New *ex ante* re-assessments of the conformity will be needed in case of substantial modifications to the AI systems.

As regards stand-alone high-risk AI systems, which are currently not covered by product safety legislation, a new compliance and enforcement mechanism is established along the lines of existing NLF legislation. However, with the exception of remote biometric identification systems, such high-risk AI systems are only subject to self-assessment of conformity by the providers. The justification provided in the explanatory notes[74] is that the combination with strong *ex post* enforcement would be an effective and reasonable solution, given the early phase of the regulatory intervention and the fact the AI sector is very innovative and expertise for auditing is only now being accumulated.[75]

[73] Wendehorst, 'The Proposal for an AIA from a Consumer Policy Perspective' (n 22) 27; C Wendehorst and Y Duller, 'Biometric Recognition and Behavioral Detection' (*European Parlament*, 2021), 63; C Wendehorst and J Hirtenlehner, 'Outlook on the future regulatory requirements for AI in Europe' (2022), 35.

[74] COM (2021) 206 final, explanatory note no 64.

[75] Critical T Schmidt and S Voeneky, Chapter 8, in this volume.

V. THE EMERGING LANDSCAPE OF AI LIABILITY LEGISLATION

While Commission proposals on AI liability, which were initially planned for the first quarter of 2022, have meanwhile been postponed to the third quarter of 2022, a draft Regulation by the European Parliament has been on the table since October 2020.[76] It was prepared in parallel with the Commission's White Paper on AI and the preparatory work for the AIA Proposal and has clearly been influenced by work at Commission level.

1. *The European Parliament's Proposal for a Regulation on AI Liability*

The cornerstone of the EP Proposal for the regulation of AI liability is a strict liability regime for the operators of 'high-risk' AI systems enumeratively listed in an Annex, accompanied by an enhanced regime of fault liability for the operators of other AI systems.

a. Strict Operator Liability for High-Risk AI Systems

According to Article 4 of the EP Proposal, operators of AI systems shall be strictly liable for any harm or damage that was caused by a physical or virtual activity, device, or process driven by an AI system. The EP Proposal ultimately adopted the division into 'frontend operator' (i.e. the person deploying the AI system) and 'backend operator' (i.e. the person that continuously controls safety-relevant features of the AI system, such as by providing updates or cloud services) that had been developed by the author of this paper and included in the 2019 EG-NTF report.[77] According to the final version of the EP Proposal, not only the frontend operator, but also the backend operator may become strictly liable. However, the backend operator's liability is covered only if it is not already covered by the PLD.[78] The only defence available to the operator is *force majeure*.[79] For the AI systems subject to strict liability, mandatory insurance is being proposed.[80]

'High-risk' AI systems for the purpose of the proposed Regulation are to be exhaustively listed in an Annex. Interestingly, the final version of the Proposal was published with the Annex left blank. The Annex attached to the first published draft from April 2020 had met with heavy resistance due to its many inconsistencies, and it may have proved too difficult to agree on a better version. Also, it seemed opportune to wait for the list of 'high-risk' AI applications that would be attached to the AIA. In any case, given the rapid technological developments and the required technical expertise, the idea is that the Commission should review the Annex without undue delay, but at least every six months, and if necessary, amend it through a delegated act.[81]

b. Enhanced Fault Liability for Other AI Systems

The EP Proposal does not only include a strict liability regime for 'high-risk' applications, but also a harmonised regime of rather strictish fault liability for all other AI systems. Article 8 provides for fault-based liability for 'any harm or damage that was caused by a physical or virtual activity, device or process driven by the AI-system', and fault is presumed (i.e. it is for the operator to show that the harm or damage was caused without his or her fault).[82] In doing so, the

[76] EP Resolution on a Civil Liability Regime for AI (n 15).
[77] NTF Expert Group (n 10) Key Findings nos 10 and 11.
[78] See Article 3(e).
[79] See Article 4(3).
[80] *Cf.* EP Resolution on a Civil Liability Regime for AI (n 15) Article 4(4).
[81] EP Resolution on a Civil Liability Regime for AI (n 15) Recommendation to the Commission no 16.
[82] In fact, the drafting is not very clear with regard to this point. Recital 17 seems to underline that fault is always presumed and that the operators need to exonerate themselves. However, Recital 19 also refers to proof of fault by the victim.

operator may rely on either of the following grounds: The first ground is that the AI-system was activated without his or her knowledge while all reasonable and necessary measures to avoid such activation outside of the operator's control were taken. The second ground is that due diligence was observed by performing all the following actions: selecting a suitable AI-system for the right task and skills, putting the AI-system duly into operation, monitoring the activities, and maintaining the operational reliability by regularly installing all available updates. It looks as if these two grounds are the only grounds by means of which operators can exonerate themselves, but Recital 18 also allows for a different interpretation, namely, that the two options listed in Article 8(2) should just facilitate exoneration by establishing 'counter-presumptions'.

The proposed fault liability regime is problematic not only because of the lack of clarity in drafting, but also because Article 8(2)(b) might be unreasonably strict, as it seems that the operator must demonstrate due diligence in all aspects mentioned, even if it is clear that lack of an update cannot have caused the damage. More importantly, in the absence of any restriction to professional operators, even consumers would face this type of enhanced liability for any kind of AI device, from a smart lawnmower to a smart kitchen stove. This would mean burdening consumers with obligations to ensure that updates are properly installed, irrespective of their concrete digital skills, and possibly confronting them with liability risks they would hardly ever have had to bear under national legal systems.

c. Liability for Physical and Certain Immaterial Harm

Article 2(1) of the Proposal declares the proposed Regulation to apply where an AI system has caused 'harm or damage to the life, health, physical integrity of a natural person, to the property of a natural or legal person or has caused significant immaterial harm resulting in a verifiable economic loss'. Article 3(i) provides for a corresponding definition of 'harm or damage'. While life, health, physical integrity, and property were clearly to be expected in such a legislative framework, the inclusion of 'significant immaterial harm resulting in a verifiable economic loss' came as a surprise. If immaterial harm or the economic consequences resulting from it – such as loss of earnings due to stress and anxiety that do not qualify as a recognised illness – is compensated through a strict liability regime whose only threshold is causation,[83] the situations where compensation is due are potentially endless and difficult to cover by way of insurance.[84]

This is so because there is no general duty not to cause significant immaterial harm of any kind to others, unless it is caused by way of non-compliant conduct (such as by infringing the law or by intentionally acting in a way that is incompatible with public policy). For instance, where AI used for recruitment procedures leads to a recommendation not to employ a particular candidate, and if that candidate, therefore, suffers economic loss by not receiving the job offer, full compensation under the EP Proposal for a Regulation would be due even if the recommendation was absolutely well-founded and if there was no discrimination or other objectionable element involved. While some passages of the report seem to choose somewhat more cautious formulations, calling upon the Commission to conduct further research,[85] Recital 16 explains very firmly that 'significant immaterial harm' should be understood as meaning harm as a result of which the affected person suffers considerable detriment, an objective and demonstrable impairment of his or her personal interests and an economic loss calculated having regard, for example, to annual average figures of past revenues and other relevant circumstances.

[83] EP Resolution on a Civil Liability Regime for AI (n 15) Article 4(1).

[84] *Cf.* T Schmidt and S Voeneky, Chapter 8, in this volume, who suggest that companies that develop or produce high-risk AI should contribute to a fund that covers damages caused by AI-driven high-risk products or services.

[85] EP Resolution on a Civil Liability Regime for AI (n 15) Recommendation to the Commission no 19.

2. *Can the EP Proposal be Linked to the AIA Proposal?*

The 2020 White Paper on AI, the EP's 2020 Proposal for an AI Liability Regulation, and the 2021 Commission Proposals for an AIA and for a new Machinery Regulation clearly have a number of parallels. They range from some identical terminology (e.g. 'AI system', 'high-risk') to the legislative technique of exhaustively listing 'high-risk' AI systems in an Annex, combined with the option for the European Commission to amend the Annex in a rather flexible procedure through delegated acts. So the question arises whether it would be possible to link an AI liability regime along the lines of the EP Proposal with the AIA Proposal in a way that the legal requirements and obligations perspective matches the liability perspective.

a. Can an AI Liability Regulation Refer to the AIA List of 'High-Risk' Systems?

The first question that arises is whether the list of 'high-risk' AI systems in the AI Liability Regulation can be identical to the list of 'high-risk' AI systems under the AIA. However, as tempting as it may be to simply refer to the AIA, it would lead to overreaching and inappropriate results. The justification for imposing strict liability that the relevant product or activity leads to significant and/or frequent harm despite the absence of any fault or any identifiable defect, mal-performance, or non-compliance does not coincide with the justification for imposing particular precautionary measures against unsafe products. While the AI systems for which strict liability is justified will most likely be a subset of the AI systems for which enhanced safety measures are justified, by far not all AI systems of the latter type should be included in a strict liability regime, for example, when they are normally safe except when clearly defective. This is underlined by the fact that the relevant players are not identical. While safety requirements are primarily addressed at the level of producers ('providers' in the AIA terminology), the EP Proposal suggests imposing strict AI liability primarily on the frontend operators ('users' in the AIA terminology), but also on the backend operators (a concept missing in the AIA). So even if something along the lines of the EP Proposal became the law it would be imperative to draft a liability-specific Annex defining 'high-risk' AI systems specifically for liability purposes. This could, for example, include big AI-driven cleaning or lawnmower robots used in public spaces, but not a small vacuum cleaner or toy robot.

b. Can the AIA Keep Liability for Immaterial Harm within Reasonable Boundaries?

As concerns fundamental rights risks, the current approach taken by the EP Proposal, which considers strict liability (alongside fault liability) for 'significant immaterial harm that results in a verifiable economic loss', has already been discarded earlier in this chapter[86] because of its failure to keep liability within any reasonable boundaries. However, the question arises whether the AIA Proposal can now assist in solving this problem.

One way of attaching liability immediately to the AIA Proposal seems to be attaching liability to the engagement in any prohibited AI practice within the meaning of Title II of the AIA Proposal, which could lead to the compensation of both material and immaterial harm thereby caused. This would be a model of non-compliance liability and fit easily into existing non-discrimination, data protection, and consumer protection legislation, all of which provide for liability for damages where harm has been caused by the engagement in prohibited practices.

Another option would be to restrict liability for immaterial harm to cases of non-conformity with the legal requirements in Title III Chapter 2 of the AIA. For instance, where training, validation, or testing data for recruitment AI fail to be relevant, representative, free of errors, and complete, as

[86] See V 1(c).

required by Article 10(4) of the AIA Proposal, the provider could be liable if an applicant was falsely filtered out by the system despite being objectively better qualified. However, it soon transpires that the legal requirements included in Title III Chapter 2 of the AIA Proposal are not optimally suited as a basis for defect liability. For many of the requirements are not so much ends in themselves that would automatically mean an AI system violates fundamental rights. Rather, some of them resemble due diligence standards that must be met during AI development, either as a quality-enhancing measure (e.g. data governance) or to facilitate monitoring (e.g. record-keeping). Non-conformity with such requirements could, therefore, justify a shift of the burden of proof, but should not in itself trigger liability. Thus, in the case of the recruitment AI system, non-conformity of training data with Article 10 should not lead to a final determination of liability but rather to the presumption that the resulting AI was defective.

VI. POSSIBLE PILLARS OF FUTURE AI LIABILITY LAW

If the AIA Proposal as it currently stands is not optimally suited for functioning as a 'backbone' for AI liability, this does not mean that the AIA as such cannot fulfil this function. Upon a closer look, not much would have to be changed in the AIA to make it an appropriate basis for future legal regimes on AI liability. At the end of the day, liability for damages caused by AI systems may have to rest on different pillars, all of which would have to rely on, or at least be aligned with, provisions in the AIA and further product safety and other law.

1. *Product Liability for AI*

The first obvious link between the AIA (and other product safety law) on the one hand and liability law on the other could be established within product liability law, which relies on the PLD. Meanwhile, it is widely accepted that the PLD must in any case be adapted to the challenges of digital ecosystems at large.[87]

a. Traditional Safety Risks

With regard to the reform of the PLD, the debate has so far been focused entirely on safety risks. Already with regard to these risks, the PLD as it currently stands is not fit to meet the challenges posed by digitalisation, not least in the light of uncertainties with regard to its scope (e.g. concerning self-standing software, including AI) and its focus on the point in time when a product is put into circulation, which fails to take into account updates, data feeds, and machine learning.[88] Where AI is involved, a victim may face particular difficulties showing that the AI system was defective. This is why no defect of the AI should have to be established by the victim for AI-specific harm caused by AI-driven products. Rather, it should be sufficient for the victim to prove that the harm was caused by an incident that might have something specifically to do with the AI (e.g. the cleaning robot making a sudden move in the direction of the victim) as contrasted with other incidents (e.g. the victim stumbling over the powered-off cleaning robot).[89]

[87] Among the plethora of pleas made in this direction, see only C Twigg-Flesner in European Law Institute (ELI) (ed), *Guiding Principles for Updating the Product Liability Directive for the Digital Age* (2021) https://europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Guiding_Principles_for_Updating_the_PLD_for_the_Digital_Age.pdf.

[88] Wendehorst and Duller, 'Safety and Liability' (n 16) 68; Koch, 'Product Liability 2.0 – Mere Update or New Version?' (n 14) 102.

[89] Wendehorst and Duller, 'Safety and Liability' (n 16) 6, 93.

b. Product Liability for Products Falling Short of 'Fundamental Rights Safety'?

As has been pointed out, the AIA Proposal also addresses fundamental rights risks. This raises the question whether also product liability might, in the future, include liability for products with a 'fundamental rights defect' or falling short of 'fundamental rights safety'.

The legal requirements described in Title III Chapter 2 of the AIA Proposal address some cloudy notion of 'adverse impact on the fundamental rights' of persons, including non-discrimination and gender equality, data protection and privacy, and the rights of the child. However, they fail to state – either in a positive or in a negative manner – what exactly the legal requirements are designed to achieve or to prevent. It is rather obvious that discrimination as far as prohibited by EU non-discrimination law, or data processing as far as prohibited by EU data protection law, is among the core effects to be prevented. However, given the much more 'fuzzy' nature of fundamental rights risks as compared with traditional safety risks, and given that there is a floating spectrum of beneficial or adverse impact on a broad variety of different fundamental rights, it is very difficult to impose liability for the materialisation of fundamental rights risks as such.

In order to achieve liability for the materialisation of fundamental rights risks as such, the first step must be to formulate an equivalent to the established concept of 'safety' in traditional product safety legislation. As far as traditional safety risks are concerned, it is possible for Article 6 (1) of the PLD to simply state: 'A product is defective when it does not provide the safety which a person is entitled to expect, taking all circumstances into account [. . .]', implicitly referring to the bulk of existing product safety law that is designed to protect 'the safety and health of persons' and similar traditional notions of safety. A corresponding concept of 'fundamental rights safety' could theoretically be derived from the AIA, in particular from the requirements for high-risk AI systems listed in Chapter 2 of Title III of the current proposal. However, in order to make these requirements operational for purposes of liability law they would have to be divided into two groups. Requirements which constitute 'AI-specific safety' (which would, by and large, be the requirements listed in Articles 13 through 15 of the draft AIA) would have to be seen as clearly separated from the requirements that are about managing safety (mostly Article 9), increasing the likelihood of safety (selected aspects of which are listed in Article 10), or documenting safety (Articles 11 and 12). Shortcomings in the technical documentation or in logging capabilities, for instance, should not be seen as a lack of 'fundamental rights safety' as such, but should rather trigger proof-related consequences in the liability context. Where technical documentation or logging capabilities are missing, or where the producer withholds logging data that would be available and potentially relevant, there could be a presumption that the missing information would have been to the detriment of the producer. Where, on the other hand, an AI system is not as accurate and robust as stated in its description or as could reasonably be expected from an AI system of the relevant kind, and therefore harm occurs (e.g. recruitment software assessing candidates has a strong gender bias and therefore female applicants are discriminated against), this lack of accuracy or robustness might trigger liability of the provider under an extended scheme of product liability. Designing such an extended scheme of product liability would, without doubt, remain to be challenging.

2. *Strict Operator Liability for 'High-Physical-Risk' Devices*

As far as death, personal injury, or property damage caused by a 'high-risk' product that includes AI for safety-relevant functions is concerned, strict liability seems to be a proper response. Again, the question arises whether the AIA can be made operational for the purposes of liability law.

a. Why AI Liability Law Needs to be More Selective than AI Safety Law

As has already been pointed out,[90] not every product that qualifies as a 'high-risk' product under the AIA fulfils the requirements that should be met for justifying strict liability (and the accompanying burden of insurance). For instance, a small robot vacuum cleaner may, under the future Machinery Regulation (if the current draft were enacted as is), be automatically classified as 'high-risk' and be subject to third party conformity assessment. It would, therefore, at least if the AI component fulfils a safety function, automatically be classified as a 'high-risk' AI system also under AIA. Similarly, a toy robot vehicle for children using AI for a safety function would be qualified as 'high-risk' under the AIA in cases where that toy is subject to third party conformity assessment,[91] (e.g. in any case where no harmonised standards exist that cover all safety requirements, or the producer has deviated from the standard).[92]

However, it would arguably be exaggerated to impose strict liability for harm caused by small toy robots or robot vacuum cleaners, in particular if that strict liability is imposed on operators. Those machines hardly ever cause significant physical harm by themselves, and if they do, it is usually because it was improper for the (frontend) operator to deploy them in the particular situation, such as where the operator of a retirement home uses an unsupervised cleaning robot in places and at times when elderly residents might stumble over it. Another possibility is that the machine is defective, for example, the vacuum cleaner, which is normally only used during the night in areas that are locked for residents, suddenly breaks loose and starts hovering when elderly residents are leaving the dining room. The problem is not so much that it would be inappropriate in the case of the retirement home to make its operator strictly liable for damage caused by the cleaning robot. Rather, the problem is that if all operators of small vacuum cleaner robots (including the millions of businesses that use them for cleaning their office space during the night, or even consumers) had to face strict liability and had to take out corresponding insurance, this would be extremely inefficient and benefit no one but the insurance industry.

b. Differentiating 'High-Risk' and 'High-Physical-Risk-As-Such'

The AIA could, therefore, be made fully operational as a 'backbone' to AI liability law if its Article 6 with Annex II drew a distinction between AI systems that are – for whatever inner logic the relevant sectoral NLF product safety legislation may follow – subject to third party conformity assessment, and AI systems that create a high physical risk as such. Needless to say, the two groups would not be mutually exclusive, as AI systems that create a high physical risk as such will often be subject to third party conformity assessments under the relevant product safety law. On the other hand, it will often be AI systems governed by 'old approach' legislation[93] that pose a high physical risk to the safety of persons as such. This means that the AIA could provide a better basis for AI liability law if these two groups of AI systems could be separated and better differentiated, either by way of restructuring and slightly redrafting Article 6 and Annex II or by drawing that distinction in a separate legal instrument on AI liability.

[90] See sub V 2(a).

[91] Article 19(3) of Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys [2009] OJ L 170/1 1, last amended by Commission Directive (EU) 2018/725 of 16 May 2018.

[92] As set out in Article 10 and Annex II of Directive 2009/48/EC. Note that the requirements are so far focused on mechanical/physical properties (e.g. sharp edges and weight), flammability, chemicals, and heavy metals restrictions, so there will be only very few AI-driven toys qualifying as 'high-risk' under the AIA.

[93] As listed in section B of Annex II and largely exempt from the AIA itself by Article 2(2) COM (2021) 202 final.

c. Avoiding Inconsistencies with Regard to Human-Driven Devices

However, it should also be borne in mind that strict liability for physical risks caused by AI-driven devices might create significant inconsistencies if not accompanied by strict liability for the same type of devices where those devices are not AI-driven but steered by humans or by technology other than AI. A victim run over by a vehicle does not care that much whether the vehicle was AI-driven or not. So if strict liability is found to be appropriate for a particular type of device of a certain minimum weight running at a certain minimum speed in public spaces (or other spaces where they typically get into contact with persons involved with the operation), this will normally be the case irrespective of whether the device is human-driven or AI-driven. For instance, large cleaning machines, lawnmowers, or delivery vehicles in public spaces might generally have to be included in strict liability regimes even where, in the relevant jurisdiction, this is so far not the case. So a strict liability regime should, at the end of the day, not be restricted to AI systems.

3. *Vicarious Operator Liability*

Vicarious liability in the sense of liability for the acts and omissions of others, such as (human) auxiliaries, might be yet another pillar of future AI liability.

a. The 'Accountability Gap' that Exists in a Variety of Contexts

Part of the problem with existing liability regimes in Member States is associated with the absence, in most legal systems, of vicarious liability for the mal-functioning of machines. Where a human cleaner knocks over a person passing by, or where a human bank clerk miscalculates a customer's credit score, there is usually fault liability of either the human auxiliary that was acting, or their employer, or both. Where, however, the person passing by is knocked over by a cleaning robot, or the credit score miscalculated by credit scoring AI, it is well possible that no one is liable at all. The AI system itself cannot be liable, but its operator may not be liable either if that operator can demonstrate that they have bought the AI system from a recognised provider and complied with all monitoring and similar duties. The producer will often not be liable as a defect in the AI system is sometimes difficult to prove, and in any case product liability (unless it will be significantly extended) only covers personal injury and property damage.

Vicarious liability would be a solution, but the rules on liability for acts or omissions of others differ vastly across the Member States and some courts insist that this kind of liability remains restricted to human auxiliaries.[94] Due to the fact that the application of vicarious liability, either directly or by analogy, is uncertain, an 'accountability gap' may exist, as very harmful activities could be conducted without anyone taking responsibility. This concerns both contexts where fault liability would normally apply and contexts where there would be non-compliance liability, and possibly other contexts.

b. Statutory or Contractual Duty on the Part of the Principal

Vicarious AI liability can only go as far as the operator of the AI would itself be liable, under national law, for violation of the same standard of conduct. This means that there must exist some statutory or contractual duty, in particular a duty of diligence, on the part of the operator. Such duties may exist in a variety of contexts, from professional care to recruitment to credit scoring to pricing, and vicarious liability may become relevant for a variety of legal frameworks, from traditional areas of tort law to non-discrimination law to data protection law to consumer and competition law.

[94] NTF Expert Group (n 10) 24 *et seq*.

Such duties could also follow from the AIA. It is, in particular, the engagement in prohibited AI practices that should lead to liability, irrespective of whether the operator was acting intentionally or negligently with regard to the fact that, for example, the AI was exploiting age-specific vulnerabilities. With an associated liability scheme in mind, it becomes even more apparent, though, that the very 'pointillistic' style of Title II of the AIA Proposal is a problem and that, if fundamental rights protection is taken seriously, it would have been necessary to have a more complete list of blacklisted AI practices plus ideally a general clause to cover unforeseen cases.

c. A Harmonised Regime of Vicarious Liability

A new European scheme of vicarious liability might restrict itself to ensuring that a principal that employs AI for a sophisticated task faces the same liability under existing Member State law as a principal that employs a human auxiliary.[95] For example, a professional user of an AI system would be liable for harm caused by any lack of accuracy or other shortcomings in the operation of the system to the same extent as that user would be liable (under the applicable national law) for the acts or omissions of a human employee mandated with the same task as the AI system. Where a human would not have been able to fulfil the same task, such as where the task requires computing capabilities exceeding those of humans, the point of reference for determining the required level of performance would be available comparable technology which the user could be expected to use.[96]

However, the EU legislator could also go one step further and introduce a fully harmonised concept of vicarious liability that does not suffer from the outset from the shortcomings we see in existing national concepts. By and large, this new European scheme of vicarious liability could provide that a business or public authority is liable for damage caused by its human auxiliaries acting within the scope of their functions, or any AI employed by the business or public authority, where these auxiliaries or AI fail to perform – for whatever reason – at the standard that could reasonably be expected from them.[97] This comes close to strict liability insofar as it requires neither fault nor a defect (or general lack of reliability in the case of human auxiliaries), but some output that does not meet the standards of conduct to be expected from a business or public authority in the fulfilment of their functions. What this level of quality is, depends on the task to be fulfilled. For instance, if it is about assessing the creditworthiness of a customer seeking credit, it would be the duty to provide proper assessment along the lines of any criteria prescribed by the law or stated by the business, and if it is about assessing candidates for a vacant position, it is again about assessing them properly, without any prohibited discrimination and duly taking into account the qualifications required for the position. Vicarious liability would, in any case, cover both safety risks and fundamental rights risks.

4. *Non-Compliance and Fault Liability*

Last but certainly not least, non-compliance and fault liability can also play an important role in the future landscape of liability for AI. In very much the same manner as Article 82 of the GDPR provides for liability of a controller or processor where that controller or processor violates their obligations under the GDPR, there could be liability under the AIA, or in a separate piece of legislation, where a provider, user or other economic operator covered by the AIA fails to comply with relevant AIA provisions, thereby causing relevant harm. This non-compliance liability

[95] Wendehorst and Duller, 'Safety and Liability' (n 16) 92.
[96] NTF Expert Group (n 10) Key Findings nos 18 and 19.
[97] This would amount to a combination between Article 6:102 (Liability for Auxiliaries) and Article 4:202 (Enterprise Liability) PETL.

might complement general fault liabiity that would continue to co-exist as a general baseline for extra-contractual liability. A breach of a duty of care that would constitute negligence could include deploying AI for a task it was not designed for, failing to provide for appropriate human oversight and other safeguards or failing to provide for necessary long-term monitoring and maintenance. Non-compliance liability and fault liability could also be merged, such as by alleviating the burden of proof for the victim under fault liability, or even reversing that burden, where obligations under the AIA have failed to be complied with.

VII. CONCLUSIONS

The potential risks associated with AI appear as normally falling into either of two dimensions: (a) 'safety risks' (i.e. death, personal injury, damage to property etc.) caused by unsafe products and activities involving AI and (b) 'fundamental rights risks' (i.e. discrimination, total surveillance, manipulation, exploitation, etc.), including risks for society at large, caused by inappropriate decisions made with the help of AI or otherwise inappropriate deployment of AI. While safety risks are highly relevant also in the AI context, fundamental rights risks are much more AI-specific.

Existing extra-contractual liability regimes can essentially be divided into four categories: fault liability, non-compliance liability, defect or mal-performance liability, and strict liability in the narrower sense. Vicarious liability can normally also be analysed as falling into one of these categories. Three out of the four categories of liability regimes are either restricted to, or heavily focused on, traditional safety risks such as death, personal injury, or property damage. It is only non-compliance liability, such as can be found in the GDPR or as an annex to EU non-discrimination law or consumer protection law, that frequently addresses also harm resulting from fundamental rights risks. Despite the fact that fundamental rights risks are more AI specific, liability for such risks seems to be largely unchartered territory, and the debate around liability for AI has largely been restricted to safety risks.

At the level of AI safety law, fundamental rights risks are now being addressed by way of prohibiting certain AI practices and by imposing mandatory legal requirements for other 'high-risk' AI systems, such as concerning data and data governance, transparency, and human oversight. While it is not impossible to use the emerging AI safety regime as a 'backbone' for the future AI liability regime, the AIA proposal, as it currently stands, is not optimally suited to help address liability for fundamental rights risks.

The future AI liability law could rest on several different pillars, such as: (a) a revised regime of product liability, which might even include liability for lack of 'fundamental rights safety'; (b) strict operator liability for death, personal injury, property damage, and possibly further safety risks caused by 'high-physical-risk' devices; (c) vicarious operator liability for mal-performance of functions carried out in the course of business activities or activities of a public authority; and (d) fault and/or non-compliance liability for the operator's own negligence and/or failure to comply with obligations following from, in particular, the AIA.

While it would be desirable to have an AI safety regime that allows an AI liability regime to dock on, it becomes apparent that the AIA Proposal has, regrettably, not been drafted with liability law in mind. Further negotiations about the AIA Proposal and the preparatory work on a future AI liability regime as well as on a potential revision of the PLD should, for the sake of consistency of Union law and of legal certainty, be more closely aligned.

13

Forward to the Past

A Critical Evaluation of the European Approach to Artificial Intelligence in Private International Law

Jan von Hein

I. INTRODUCTION

On 2 October 1997, the Member States of the European Union (EU) signed the Treaty of Amsterdam and endowed the European legislature with a competence in the field of private international law that is now found in Article 81(2)(c) of the Treaty on the Functioning of the European Union.[1] In the following two decades, the EU created an expanding body of private international law.[2] In particular, the Rome II Regulation on the law applicable to non-contractual obligations was enacted on 11 July 2007.[3] Only eleven months later, the Rome I Regulation on the law applicable to contractual obligations was adopted.[4] Although both Regulations are already rather comprehensive, gaps as well as inconsistencies remain.[5] In light of the rapid technological development since 2009, the issue as to whether there is a need for specific rules on the private international law of artificial intelligence (AI) has to be addressed.[6] After the European Parliament's JURI Committee had presented a proposal for a civil liability

[1] Article 61(c) in conjunction with Article 65(b) of the Treaty of Amsterdam [1997] OJ C340/173 establishing the European Community; today Article 81(1) and (2)(c) of the Treaty on the Functioning of the European Union [2012] OJ C326/01; for an early assessment, see J Basedow, 'The Communitarization of the Conflict of Laws under the Treaty of Amsterdam' (2000) 37 *CML Rev* 687; on more recent developments, J von Hein, 'EU Competence to Legislate in the Area of Private International Law and Law Reforms at the EU Level' in P Beaumont and others (eds), *Cross-Border Litigation in Europe* (2017) 19.

[2] See G Rühl and J von Hein, 'Towards a European Code on Private International Law?' (2015) 79 *RabelsZ* 701 *et seq*. (hereafter Rühl and von Hein, 'Towards a European Code').

[3] Regulation (EC) 864/2007 of the European Parliament and of the Council of 11 July 2007 on the law applicable to non-contractual obligations (Rome II), [2007] OJ L 199/40; on the legislative history up to 2003, see J von Hein, 'Die Kodifikation des europäischen Internationalen Deliktsrechts' (2003) 102 *ZVglRWiss* 528, 529–533; up to 2007, J von Hein, 'Die Kodifikation des europäischen IPR der außervertraglichen Schuldverhältnisse vor dem Abschluss?' (2007) *Versicherungsrecht* 440; on the final compromise between the Council and the Parliament, see R Wagner, 'Das Vermittlungsverfahren zur Rom II-VO' in D Baetge, J von Hein, and M von Hinden (eds), *Die richtige Ordnung, Festschrift für Jan Kropholler* (2008) 715 (hereafter Wagner, 'Das Vermittlungsverfahren').

[4] Regulation (EC) No 593/2008 of the European Parliament and of the Council of 17 June 2008 on the law applicable to contractual obligations (Rome I), 2008 OJ L 177/6.

[5] Rühl and von Hein, 'Towards a European Code' (n 2) 713–715.

[6] For a general survey, see L Wetenkamp, 'IPR und Digitalisierung: Braucht das Internationale Privatrecht ein Update?' (*Beiträge zum Transnationalen Wirtschaftsrecht Volume 161*, April 2019) https://telc.jura.uni-halle.de/sites/default/files/BeitraegeTWR/Heft161.pdf (hereafter Wetenkamp, 'IPR und Digitalisierung'); on autonomous driving in particular, see T Kadner Graziano, 'Cross-Border Traffic Accidents in the EU: The Potential Impact of Driverless Cars, Study for the JURI Committee' (*European Parliament*, June 2016) www.europarl.europa.eu/RegData/etudes/STUD/2016/571362/IPOL_STU(2016)571362_EN.pdf (hereafter Kadner Graziano, 'Driverless Cars').

regime for AI in April 2020,[7] the European Parliament adopted – with a large margin – a pertinent resolution with recommendations to the Commission on 20 October 2020.[8] This resolution is part of a larger regulatory package on issues of AI.[9] The draft regulation (DR) proposed in this resolution is noteworthy not only with regard to the rules on substantive law that it contains,[10] but also from a choice-of-law perspective because it introduces new, specific conflicts rules for AI-related aspects of civil liability.[11] In the following contribution, I analyse and evaluate the European Parliament's proposal against the background of the already existing European regulatory framework on private international law, in particular the Rome I and II Regulations.

II. THE CURRENT EUROPEAN FRAMEWORK

1. *The Goals of PIL Harmonisation*

The basic economic rationale underlying the Rome II Regulation is succinctly captured in its Recital 6, which reads as follows:

> The proper functioning of the internal market creates a need, in order to improve the predictability of the outcome of litigation, certainty as to the law applicable and the free movement of judgments, for the conflict-of-law rules in the Member States to designate the same national law irrespective of the country of the court in which an action is brought.

This Recital epitomises the basic tenet of the methodology developed by *Friedrich Carl von Savigny* in the nineteenth century, in other words, the goal of international decisional harmony.[12] The Commission's explanation for its Rome II draft of 2003 is even more explicit with regard to the deterrence of forum shopping: unless conflicts rules for non-contractual obligations become unified, '[t]he risk is that parties will opt for the courts of one Member State rather than another simply because the law applicable in the courts of this State would be more favourable to them.'[13] The explanation for the draft of 2003 also makes clear that a unification of tort conflicts rests on a sound economic rationale, the reduction of transaction costs borne by the parties. A European Regulation on tort conflicts 'allows the parties to confine themselves to

[7] European Parliament, Draft Report 2020/2014(INL) (*European Parliament*, 27 April 2020) www.europarl.europa.eu/doceo/document/JURI-PR-650556_EN.pdf.

[8] The text of this resolution is available at www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.pdf.

[9] For an overview, see the Parliament's press release 'Parliament Leads the Way on First Set of EU Rules for Artificial Intelligence' (European Parliament, 20 October 2020) www.europarl.europa.eu/news/en/press-room/20201016IPR89544/parliament-leads-the-way-on-first-set-of-eu-rules-for-artificial-intelligence; on subsequent developments, see the overview by A Pato, 'The EU's Upcoming Regulatory Framework on Artificial Intelligence and Its Impact on PIL' (*EAPIL Blog*, 12 July 2021) https://eapil.org/2021/07/12/the-eus-upcoming-regulatory-framework-on-artificial-intelligence-and-its-impact-on-pil.

[10] On those rules, see H Sousa Antunes, 'Civil Liability Applicable to Artificial Intelligence: A Preliminary Critique of the European Parliament Resolution of 2020' (*SSRN*, 8 January 2021) https://ssrn.com/abstract=3743242; G Wagner, 'Haftung für Künstliche Intelligenz: Eine Gesetzesinitiative des Europäischen Parlaments' (2021) 29 *ZEuP* 545.

[11] On the general issues of AI and private international law, see Wetenkamp, 'IPR und Digitalisierung' (n 6); see also the conference report by S Arnold, T Eick, and C Hornung, 'Conference Report: Conflict of Laws 4.0 (Münster, Germany)' (*Conflict of Laws*, 14 January 2020) https://conflictoflaws.net/2020/conference-report-conflict-of-laws-4-0-munster-germany.

[12] FC von Savigny, *A Treatise on the Conflict of Laws* (1880) 69 et seq.

[13] Commission's proposal for a regulation of the European Parliament and of the Council on the law applicable to non-contractual obligations (Rome II), COM(2003) 427 final, reprinted in J Ahern and W Binchy (eds), *The Rome II Regulation on the Law Applicable to Non-Contractual Obligations* (2009) 301, 303.

studying a single set of conflict rules, thus reducing the cost of litigation and boosting the foreseeability of solutions and certainty as to the law.'[14] This rationale is particularly important for tort conflicts, because, contrary to contract conflicts, a choice of the applicable law *ex ante* was traditionally not available in many jurisdictions.[15] Even if the parties enjoy that possibility, they will frequently not be able to exercise this right because they do not anticipate an accident to happen.[16] Accordingly, clear objective conflicts rules have significantly greater weight in tort than in contract cases.[17] This is an important factor facilitating the emergence of new technologies with cross-border implications, such as driverless cars.[18]

Moreover, the force of a practical example that would emanate from a successful codification of European conflicts rules on AI must not be underestimated. Although the initial American reaction towards the Rome II Regulation was rather critical, denouncing the final text as a 'missed opportunity' to transplant US doctrines to Europe,[19] there is a palpable transatlantic interest in recent European developments and the lessons that these may hold for the United States.[20] A well-known American conflicts scholar even recommended the European codification of tort conflicts as a model for further US legislation.[21] While the 'end of history' for private international law (i.e. a full convergence of US and European conflict of laws in torts),[22] is still a long road ahead, a successful EU legislation on the law applicable to liability issues of AI will certainly increase the prospects for creating harmonised conflicts rules in this area on a global level.

2. *The Subject of Liability*

Both the Rome I and II Regulations only address the liability of natural persons[23] and 'companies and other bodies, corporate or unincorporated'.[24] Thus, the question arises as to whether an AI system could be classified as another 'unincorporated body' within the meaning of these provisions.[25] There is a parallel discussion about attributing legal personality to AI-systems in substantive private law.[26] Although the mere wording of the English version of the Rome I and II

[14] Ibid, 305.
[15] See J von Hein, 'Art 14 Rome II para 1' in GP Calliess and M Renner (eds), *Rome Regulations – Commentary* (3rd ed. 2020) with further references.
[16] G Hohloch, 'Place of Injury, Habitual Residence, Closer Connection and Substantive Scope: The Basic Principles' (2007) 9 *YbPIL* 1.
[17] Ibid, 2.
[18] Cf. Kadner Graziano, 'Driverless Cars' (n 6) 57.
[19] SC Symeonides, 'Rome II and Tort Conflicts: A Missed Opportunity' (2008) 56 *Am J Comp L* 173; but cf. the balanced evaluation by P Hay, 'Contemporary Approaches to Non-Contractual Obligations in Private International Law (Conflict of Laws) and the European Community's "Rome II" Regulation' (2007) 7 *EuLF* I-137, I-151, who calls the Rome II Regulation a 'major achievement'.
[20] Cf. SC Symeonides, 'The American Revolution and the European Evolution in Choice of Law: Reciprocal Lessons' (2008) 82 *Tul L Rev* 1741.
[21] PJ Kozyris, 'Rome II: Tort Conflicts on the Right Track! A Postscript to Symeon Symeonides "Missed Opportunity"' (2008) 56 *Am J Comp L* 471.
[22] For an earlier assessment of the perspectives for a convergence of US and European approaches to tort conflicts, see J Kropholler and J von Hein, 'From Approach to Rule-Orientation in American Tort Conflicts?' in JAR Nafziger and SC Symeonides (eds), *Law and Justice in a Multistate World: Essays in Honor of Arthur T von Mehren* (2002) 317.
[23] Rome I, Article 19 (1) 2nd sentence; Rome II, Article 23(2).
[24] Rome I, Article 19(1) 1st sentence; Rome II, Article 23(1).
[25] See Wetenkamp, 'IPR und Digitalisierung' (n 6) 16 *et seq*.
[26] See, e.g., G Teubner, 'Digitale Rechtssubjekte? Zum Privatrechtlichen Status Autonomer Softwareagenten' (2018) 218 *AcP* 155; cf. also, from the perspective of public international law, T Burri, 'International Law and Artificial Intelligence' (2017) 60 *German Yb Int'l L* 91, 95–98.

Regulations would arguably allow such an innovative interpretation, other linguistic versions suggest a narrower, more traditional reading of the Regulations (e.g. the German one, which speaks of 'Gesellschaften, Vereine und juristische Personen'). Since the law applicable to legal personality is not yet determined by EU private international law, but remains subject to domestic choice-of-law rules within the boundaries of the freedom of establishment,[27] it would be unwise to burden the Rome I and II Regulations with a regulatory aspect that is, from the point of view of international contract and tort law, merely an incidental question. Thus, the law applicable to legal personality will have to be determined by other measures, e.g. by a regulation based on the draft presented by the European Group for Private International Law in 2016.[28]

3. *Non-Contractual Obligations: The Rome II Regulation*

a. Scope

The Rome II Regulation determines the law applicable to non-contractual obligations, in particular torts. The notion of 'non-contractual obligation' must be interpreted as an autonomous concept.[29] It covers both strict and fault-based liability.[30] Generally speaking, all types of harm or damage are covered, such as physical damage to property, pure economic loss, and immaterial harm.[31] The Rome II Regulation is limited to civil and commercial matters;[32] notably, it does not cover the liability of the state for acts and omissions in the exercise of state authority.[33] Thus, the law applicable to a Member State's liability for the use of AI for the purpose of international police surveillance or military operations, for example, is determined by domestic choice-of-law rules.[34] Moreover, the Rome II Regulation is not applicable to non-contractual obligations arising out of violations of privacy and rights relating to personality, including defamation.[35] Therefore, the law applicable to any kind of use of AI that violates a person's right to privacy or causes damage to their reputation must still be determined by domestic choice-of-law rules, such as Articles 40–42 of the German EGBGB.[36] Finally, although the rules of the Rome II Regulation are of European origin, they shall be applied whether or not the law specified by them is the law of an EU Member State.[37] Thus, according to this principle of 'universal application', even if an AI system operated by a British company causes damage to a

[27] TFEU, Articles 49 and 54; see J von Hein, 'Corporations in European Private International Law: From Case-Law to Codification?' (2015) 17 *JYIL* 90.

[28] European Group for Private International Law, Draft Rules on the Law Applicable to Companies and Other Bodies, Milan (*GEDIP*, 16–18 September 2016) https://gedip-egpil.eu/wp-content/uploads/2016/09/Societe-TxtSousGroup-1.pdf; for closer analysis, see J von Hein, 'Der Vorschlag der GEDIP für eine EU-Verordnung zum Internationalen Gesellschaftsrecht' in B Hess, E Jayme, and HP Mansel (eds), *Europa als Rechts- und Lebensraum, Liber Amicorum für Christian Kohler* (2018) 551.

[29] Rome II, Recital 11 2nd sentence; on the principle of autonomous interpretation of Rome II, see CJEU, Case C-350/14 *Florin Lazar v Allianz SpA* (10 December 2015) para 21 (hereafter CJEU, *Florin Lazar*).

[30] Rome II, Recital 11 3rd sentence.

[31] Rome II, Article 2(2); CJEU, *Florin Lazar* (n 29) para 22.

[32] Rome II, Article 1(1) 1st sentence.

[33] Rome II, Article 1(1) 2nd sentence.

[34] On international governmental liability for German military operations in Afghanistan, see BGHZ 212, 173 (Bundesgerichtshof III ZR 140/15).

[35] Rome II, Article 1(2)(g).

[36] Einführungsgesetz zum Bürgerlichen Gesetzbuch (EGBGB) – Introductory Act to the Civil Code of September 21, 1994, *Federal Law Gazette* 1994 I 2494, as amended by the Gesetz zum Internationalen Güterrecht und zur Änderung von Vorschriften des Internationalen Privatrechts, *Federal Law Gazette* 2018 I 2573, 2580; English translation by J Mörsdorf-Schulte available at www.gesetze-im-internet.de/englisch_bgbeg/.

[37] Rome II, Article 3.

person in Switzerland, the court of an EU Member State will determine the law applicable to such a case pursuant to the Rome II Regulation.[38]

b. The General Rule (Article 4 Rome II)

The basic rule for torts in general is found in Article 4(1) Rome II, which refers to the place of injury. Recital 15 Rome II acknowledges that '*lex loci delicti* is the basic solution for non-contractual obligations in virtually all the Member States'. Nevertheless, the diverging interpretations of this principle by various Member States' legislatures and courts in complex cases (place of injury, place of acting, or even both under the so-called theory of ubiquity) had in the past led to considerable legal uncertainty.[39] The preference for the place of injury is justified because, generally speaking, it strikes 'a fair balance' between the interest of the person claimed to be liable to foresee the applicable law and the interests of the person sustaining the damage.[40] From an economic point of view, the place of injury will usually lead to a fair distribution of the costs for obtaining the relevant legal information: In most cases, the person claimed to be liable should be able to anticipate that his or her acts may cause harm in another country, whereas the victim should be able to rely on the legal standard of the environment to which he or she exposed his or her body or property.[41] While the tortfeasor is thus forced to internalise the costs for negative externalities arising in other countries,[42] the victim is given the opportunity to structure his or her insurance in accordance with the law to which he or she is presumably accustomed.[43] Since Article 4(1) Rome II is based on the idea of striking 'a fair balance' between the alleged tortfeasor and victim, this neutral provision must not be interpreted in a one-sided fashion that favours the plaintiff. The Rome II Regulation does not, as a general principle, embrace the plaintiff-friendly principle of ubiquity found in German or Italian private international law.[44]

[38] For further details, see A Halfmeier, 'Article 2 Rome II paras 1–8' in GP Calliess and M Renner (eds), *Rome Regulations: Commentary* (3rd ed. 2020).

[39] See Rome II, Recital 15: 'The principle of the *lex loci delicti commissi* is the basic solution for non-contractual obligations in virtually all the Member States, but the practical application of the principle where the component factors of the case are spread over several countries varies. This situation engenders uncertainty as to the law applicable'; cf. T Kadner Graziano, 'General Principles of Private International Law of Tort in Europe' in J Basedow, H Baum, and Y Nishitani (eds), *Japanese and European Private International Law in Comparative Perspective* (2008) 243, 247; A Nuyts, 'La règle générale de conflit de lois en matière non contractuelle dans le Règlement Rome II' (2008) *Rev dr comm belge* 489, 492.

[40] Rome II, Recital 16: 'Uniform rules should enhance the foreseeability of court decisions and ensure a reasonable balance between the interests of the person claimed to be liable and the person who has sustained damage. A connection with the country where the direct damage occurred (*lex loci damni*) strikes a fair balance between the interests of the person claimed to be liable and the person sustaining the damage, and also reflects the modern approach to civil liability and the development of systems of strict liability.'

[41] J von Hein, *Das Günstigkeitsprinzip im Internationalen Deliktsrecht* (1999), 217–220; K Thorn, 'Art 4 Rome II para 1' in C Grüneberg, *Bürgerliches Gesetzbuch* (81st ed. 2022).

[42] FG Alférez, 'The Rome II Regulation: On the Way Towards a European Private International Law Code' (2007) *EuLF* I-77, I-84; L de Lima Pinheiro, 'Choice of Law on Non-Contractual Obligations between Communitarization and Globalization: A First Assessment of EC Regulation Rome II' (2008) 44 *RDIPP* 5, 16.

[43] Cf. J Basedow, 'EC Conflict of Laws: A Matter of Coordination' in L de Lima Pinheiro (ed), *Seminário Internacional sobre a Comunitarizaçao do Direito Internacional Privado* (2005) 26; A Junker, 'Die Rom II-Verordnung: Neues Internationales Deliktsrecht auf europäischer Grundlage' (2007) *NJW* 3675, 3678 (noting that the place of injury will frequently coincide with the victim's habitual residence); T Petch, 'The Rome II Regulation: An Update, Part I' (2006) *JIBLR* 449, 454.

[44] Article 40(1) of the German EGBGB (n 36); Article 62(1) of the Italian Code on Private International Law of May 31, 1995, Legge n. 218, Riforma del sistema italiano di diritto internazionale privato, Supplemento ordinario n 68 alla Gazzetta Ufficiale n 128, June 3, 1995, reprinted in (1997) 61 *RabelsZ* 344 (hereafter Italian PIL Code); cf. A Junker,

The Rome II Regulation contains a significant number of specific rules for special torts.[45] This considerably reduces the weight that the general rule has to carry, which applies only 'unless otherwise provided for in this Regulation'.[46] The main group of cases of practical importance that are exclusively governed by the general rule instead of specific rules are traffic accidents.[47] However, even in this regard, the scope of application of Article 4 Rome II is limited in practice. The full communitarisation of private international law is impeded by the fact that there already exist two supranational instruments dealing with important areas of tort conflicts, namely, the Hague Convention on the law applicable to Traffic Accidents (HCTA) and the Hague Convention on the law applicable to Products Liability (HCP).[48] Both conventions count several EU Member States among their parties.[49] Those Member States were (and are) unwilling to withdraw from the respective conventions.[50] Since the EU could arguably not terminate their membership without their consent, rules governing the collision between EU conflicts rules and the Hague conventions had to be invented.[51] The solution finally codified in the Rome II Regulation provides that the Regulation does not prejudice the application of existing conventions that contain conflicts rules for non-contractual obligations.[52] The Rome II Regulation takes precedence, however, over conventions concluded exclusively between two or more of them insofar as such conventions concern matters governed by the Regulation.[53] Since both pertinent Hague conventions have a sizeable number of non-EU state parties, this exception is of little practical use.[54] Even if a traffic accident is only connected with, for example, France and Germany, French courts have to apply the HCTA, whereas a German court must determine the applicable law under the Rome II Regulation.[55] Thus, in two of the most important areas of tort conflicts, traffic accidents and product liability, European private international law remains fragmented and continues to offer ample possibilities of forum shopping.[56] This situation is exacerbated by the fact that the Rome II Regulation excludes the

'Kollisionsnorm und Sachrecht im IPR der unerlaubten Handlung' in R Michaels and D Solomon (eds), *Liber Amicorum Klaus Schurig* (2012) 81, 82 *et seq*.

[45] Rome II, Articles 5 to 9.

[46] Rome II, Article 4(1).

[47] On this group of cases, see J von Hein, 'Article 4 and Traffic Accidents' in J Ahern and W Binchy (eds), *The Rome II Regulation on the Law Applicable to Non-Contractual Obligations* (2009) 153; A Junker, 'Das Internationale Privatrecht der Straßenverkehrsunfälle nach der Rom II-Verordnung' (2008) *JZ* 169; T Kadner Graziano, 'Internationale Verkehrsunfälle' (2011) *ZVR* 40.

[48] Hague Convention on the Law Applicable to Traffic Accidents of May 4, 1971, in Hague Conference on Private International Law (ed), Statute – Conventions – Protocol – Principles, The Hague 2020, No. 19; English text also available at www.hcch.net/index_en.php?act=conventions.text&cid=81; Hague Convention on the Law Applicable to Products Liability of October 2, 1973, in Hague Conference on Private International Law (ed), Statute – Conventions – Protocol – Principles, The Hague 2020, No. 22 and (1973) 37 *RabelsZ* 594.

[49] HCTA: Austria, Belgium, Croatia, Czech Republic, France, Latvia, Lithuania, Luxembourg, The Netherlands, Poland, Slovakia, Slovenia, Spain; HCP: Croatia, Finland, France, Luxembourg, the Netherlands, Slovenia, Spain.

[50] On the negotiations, see the detailed account by Wagner, 'Das Vermittlungsverfahren' (n 3) 726 *et seq*.

[51] For a closer analysis of the problems under public international and EU law, see C Brière, 'Réflexions sur les interactions entre la proposition de règlement "Rome II" et les conventions internationales' (2005) 132 *Clunet* 677.

[52] Rome II, Article 28(1); see G Garriga, 'Relationships Between "Rome II" and Other International Instruments', (2007) 9 *YbPIL* 137.

[53] Rome II, Article 28(2).

[54] HCTA: Belarus, Bosnia & Hercegovina, Macedonia, Montenegro, Morocco, Serbia, Switzerland, Ukraine; HCP: Macedonia, Montenegro, Norway, Serbia.

[55] H Ofner, 'Die Rom II-Verordnung – Neues Internationales Privatrecht für außervertragliche Schuldverhältnisse in der Europäischen Union' (2008) *ZfRV* 2008, 1315 *et seq*.

[56] A Staudinger, 'Das Konkurrenzverhältnis zwischen dem Haager Straßenverkehrsübereinkommen und der Rom II-VO' in D Baetge, J von Hein, and M von Hinden (eds), *Die richtige Ordnung, Festschrift für Jan Kropholler* (2008),

possibility of renvoi.[57] Thus, cases involving driverless cars, for example, may be subject to different laws in various Member States.[58]

The *lex loci damni*[59] is displaced in cases where the person claimed to be liable and the person sustaining the damage both have their habitual residence in the same country at the time when the damage occurs.[60] This rule had been familiar to many European codifications already before Rome II was enacted.[61] Again, it is a legitimate expression of the basic economic rationale underlying the Regulation: '[I]n most cases the common residence rule guarantees lower litigation costs, more efficient court administration, and international harmony of decisions'.[62] Usually, parties who share a common habitual residence will litigate in the country where they live; moreover, their insurance coverage will, in most cases, be structured according to the standards prevailing in this country.[63]

Article 4(1) and (2) Rome II are coupled with an escape clause that is meant to provide for a sufficient degree of judicial discretion in the individual case.[64] The final paragraph, which is rather an open-ended standard than a rule, combines a fairly general approach in its first sentence (manifestly closer connection) with a particular example of such a connection (relationship between the parties, for example, a contract) in its second sentence. As Recital 14 Rome II shows, the drafters of the Regulation were mindful of the tension between 'the requirement of legal certainty' on the one hand and the 'need to do justice in individual cases' on the other. The Recital explains that

> this Regulation provides for a general rule but also for specific rules and, in certain provisions, for an 'escape clause' which allows a departure from these rules where it is clear from all the circumstances of the case that the tort/delict is manifestly more closely connected with another

671; T Thiede and M Kellner, '"Forum shopping" zwischen dem Haager Übereinkommen über das auf Verkehrsunfälle anzuwendende Recht und der Rom-II-Verordnung' (2007) *Versicherungsrecht* 1624.

[57] Rome II, Article 24.

[58] See in more detail Kadner Graziano, 'Driverless cars' (n 6) 37 *et seq*.

[59] Rome II, Article 4(1).

[60] Rome II, Article 4(2).

[61] For example, EGBGB, Article 40(2) (n 36); Article 2(3) Wet Conflictenrecht Onrechtmatige Daad of April 11, 2001, (2001) Staatsblad van het Koninkrijk der Nederlanden, No 190, German translation in (2004) IPRax 157, now repealed and substituted by Article 159 Book 10 of the Dutch Civil Code, 19 May 2011, (2011) Staatsblad van het Koninkrijk der Nederlanden, No 272, English translation in (2011) 13 *YbPIL* 657, which mandates an analogous application of the Rome II Regulation to cases outside of its scope; Article 99(1) No 1 Loi du 16 juillet 2004 portant le Code de droit international privé (Belgian Law of July 16, 2004, holding the Code of Private International Law), (2004) Moniteur Belge 57344 (French/Dutch), official German translation in (2005) Belgisch Staatsblad 48274, English translation in (2006) 70 *RabelsZ* 358; some codifications take citizenship into account as well, for example, Article 62(2) Italian PIL Code (n 44); Article 45(3) of the Portuguese Civil Code (Código Civil Português) Decreto-Lei (n 47) 344 of November 25, 1966, in W Riering (ed), *IPR-Gesetze in Europa* (1997) 108.

[62] T Dornis, 'When in Rome, Do as the Romans Do? A Defense of the Lex Domicilii Communis in the Rome II-Regulation' (2007) *EuLF* I-152, I-157; it is not convincing to argue that the parties could reach the same result by choosing the applicable law pursuant to Article 14 Rome II, see H Unberath, J Cziupka and S Pabst 'Article 4 Rome II para 63' in T Rauscher (ed), *Europäisches Zivilprozess- und Kollisionsrecht (EuZPR/EuIPR), Kommentar, Volume 3: Rom I-VO, Rom II-VO* (4th ed. 2016), because it will in many cases be impossible to reach a consensus on the applicable law after the accident has occurred; cf. G Rühl, 'Article 4 Rome II para 85' in B Gsell and others (eds), *Beck-Online Großkommentar* (1 December 2017).

[63] Cf. A Junker, 'Article 4 Rome II para 37' in F J Säcker and others (eds), *Münchener Kommentar zum Bürgerlichen Gesetzbuch, Volume 13: Internationales Privatrecht II* (8th ed. 2021); T Kadner Graziano, 'Le nouveau droit international privé communautaire en matière de responsabilité extracontractuelle' (2008) 97 *Rev crit dr int priv* 445, 462; C von Bar and P Mankowski, *Internationales Privatrecht Volume 2* (2nd ed. 2019) para 188.

[64] Rome II, Article 4(3).

country. This set of rules thus creates a flexible framework of conflict-of-law rules. Equally, it enables the court seised to treat individual cases in an appropriate manner.

Finally, Article 14 Rome II provides for a modern and liberal approach to party autonomy for non-contractual obligations, allowing a choice of the applicable law both *ex post* and, provided certain conditions are met, *ex ante*.[65] The reasons for this liberal approach are spelled out in the first sentence of Recital 31: 'To respect the principle of party autonomy and to enhance legal certainty, the parties should be allowed to make a choice as to the law applicable to a non-contractual obligation.' Party autonomy enhances legal certainty in two ways.[66] First, the flexible approach of the Regulation, which is characterised by a rather generous array of escape clauses,[67] introduces a potential source of litigation that must be balanced by giving parties the possibility of quickly resolving any dispute on the applicable law.[68] Secondly, the substantive laws of the Member States are characterised by significant divergences as far as the proper boundaries between tort and contract law are concerned. This is particularly true for cases such as pre-contractual liability, liability for pure economic loss, and the protection of third persons who are not a party to an existing contract with the person claimed to be liable.[69] Thus, parties who want to avoid a protracted litigation on issues of classification are well advised to choose the law applicable not only to their contractual obligations, but also to their non-contractual obligations.[70]

c. The Rule on Product Liability (Article 5 Rome II)

With regard to product liability, Article 5 Rome II strives to create a balance between an effective protection of the victim, who is often a consumer and typically regarded as the weaker party, on the one hand, and the producer's interest in foreseeability of the applicable law, on the other.[71]

Article 5(1) Rome II presupposes a damage 'caused by a product'. The notion of 'product' must be interpreted autonomously;[72] the Commission's Explanatory Memorandum of 2003[73]

[65] For a comprehensive monographic treatment, see A Vogeler, *Die freie Rechtswahl im Kollisionsrecht der außervertraglichen Schuldverhältnisse* (2013).

[66] A Briggs, *Agreements on Jurisdiction and Choice of Law* (2008) para 10.72 ('entirely rational, and a great step forward'); Editorial Comments, (2007) 44 *CML Rev* 1567, 1570 ('[L]egal certainty for the parties is the winner'); E O'Hara O'Connor and L Ribstein, 'Rules and Institutions in Developing a Law Market: Views from the United States and Europe' (2008) 82 *Tul L Rev* 2147, 2167 et seq.; but cf. also TM de Boer, 'Party Autonomy and Its Limitations in the Rome II Regulation' (2007) 9 *YbPIL* 19, 22 (criticising Recital 31 as 'not very convincing') (hereafter de Boer, 'Party Autonomy').

[67] In particular Rome II, Articles 4(3) and 5(2).

[68] A functional complementarity ignored by de Boer, 'Party Autonomy' (n 66) 22.

[69] C von Bar and U Drobnig, *Study on Property Law and Non-Contractual Liability as They Relate to Contract Law* (European Commission – Health and Consumer Protection Directorate-General, SANCO/2002/B5/010, 2004).

[70] J von Hein, 'Rechtswahlfreiheit im Internationalen Deliktsrecht' (2000) 64 *RabelsZ* 595, 601; P Picht, 'Article 14 Rome II para 18' in T Rauscher (ed), *Europäisches Zivilprozess- und Kollisionsrecht EuZPR/EuIPR Kommentar, Volume 3: Rom I-VO, Rom II-VO* (4th ed. 2016).

[71] Cf. Rome II, Recital 20.

[72] A Junker 'Article 5 Rome II para 13' in F J Säcker and others (eds), *Münchener Kommentar zum Bürgerlichen Gesetzbuch, Volume 13: Internationales Privatrecht II* (8th ed. 2021); M Lehmann, 'Article 5 Rome II para 24' in R Hüßtege and HP Mansel (eds), *Bürgerliches Gesetzbuch: Rom Verordnung, Nomos-Kommentar, Volume 6: EuErbVO, HUP* (3rd ed. 2019) (hereafter Lehmann, 'Article 5 Rome II para 24').

[73] Commission's Proposal for a Rome II Regulation (n 13), COM(2003) 427 final, 14; concurring *A Junker* 'Article 5 Rome II para 3' in FJ Säcker and others (eds), *Münchener Kommentar zum Bürgerlichen Gesetzbuch, Volume 13: Internationales Privatrecht II* (8th ed. 2021); Lehmann, 'Article 5 Rome II para 24' (n 72); K Thorn, 'Article 5 Rome II para 3' in C Grüneberg (ed), *Bürgerliches Gesetzbuch* (81st ed. 2022); H Unberath, J Cziupka, and S Pabst, 'Article 5 Rome II para 38' in T Rauscher (ed), *Europäisches Zivilprozess- und Kollisionsrecht (EuZPR/EuIPR), Kommentar, Volume 3: Rom I-VO, Rom II-VO* (4th ed. 2016); O Remien, 'Art. 5 Rome II para 4' in HT Soergel, *Bürgerliches*

refers to the definition found in the EU Directive on Product Liability.[74] The substantive EU law on product liability so far only applies to physical goods.[75] Thus, strict liability for data processing cannot be based on the current Product Liability Directive.[76] A working group hosted by the European Law Institute has recently published a paper on giving the Product Liability Directive a digital 'update', but this reform process is still in its first stages.[77] Although the rules of the current Product Liability Directive may be extended to cover standard software delivered on a DVD, for example,[78] it is controversial whether software that was designed to meet the specific needs of the customer could be classified as a 'product'.[79] Those delineations are generally transferred to Article 5(1) Rome II.[80] In cases of autonomous driving, however, the software will be sold as an integral part of a car. In cases where software is embedded in a physical good, both the Product Liability Directive and Article 5(1) Rome II apply.[81]

The cascade of connections found in Article 5 Rome II is structured as follows: first, parties may choose the law applicable to product liability claims under the general provision on party autonomy.[82] Likewise, the Rome II Regulation provides for an accessory connection of product liability claims to a pre-existing relationship, such as a contract, between the parties.[83] Both steps constitute major improvements compared to the Hague Convention on the law applicable to product liability,[84] which failed to include such rules.

Secondly, if both parties have their habitual residence in the same country, the law of that state applies.[85]

Gesetzbuch mit Einführungsgesetz und Nebengesetzen BGB, Volume 27/1: Rom II-VO; Internationales Handelsrecht; Internationales Bank- und Kapitalmarktrecht (13th ed. 2019) (hereafter Remien, 'Art. 5 Rome II para 4').

[74] Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L 210/29 (Product Liability Directive); as amended by Directive 1999/34/EC of the European Parliament and of the Council of 10 May 1999, [1999] OJ L 141/20.

[75] Product Liability Directive, Article 2 1st sentence; see G Wagner, '§ 2 ProdHaftG para 15' in FJ Säcker and others (eds), *Münchener Kommentar zum Bürgerlichen Gesetzbuch, Volume 7: Schuldrecht – Besonderer Teil IV* (8th ed. 2020) (hereafter Wagner, '§ 2 ProdHaftG para 15'); J Oechsler, '§ 2 ProdHaftG para 64' in J von Staudinger (ed) *Kommentar zum Bürgerlichen Gesetzbuch mit Einführungsgesetz und Nebengesetzen, Buch 2: Recht der Schuldverhältnisse, §§ 826–829; ProdHaftG (Unerlaubte Handlungen 2, Produkthaftung)* (2014) (hereafter Oechsler, '§ 2 ProdHaftG para 64'); on product liability in the USA, cf. Restatement (Third) of Torts: Products Liability § 19 (1998), with further references in Comment d; for closer analysis, see WC Powers Jr., 'Distinguishing Between Products and Services in Strict Liability' (1984) 62 *NCL Rev* 415, 418, 425; MD Scott, 'Tort Liability for Vendors of Insecure Software: Has the Time Finally Come?' (2008) 67 *Md L Rev* 425; FE Zollers and others, 'No More Soft Landings for Software: Liability for Defects in an Industry that Has Come of Age' (2005) 21 *Santa Clara Computer and High Tech LJ* 745.

[76] T Hoeren, 'Review of "Nils Jansen, Die Struktur des Haftungsrechts"' (2004) 121 *SavZ/Germ* 590, 593.

[77] C Twigg-Flesner (ed), *Guiding Principles for Updating the Product Liability Directive for the Digital Age* (2021), available at https://europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Guiding_Principles_for_Updating_the_PLD_for_the_Digital_Age.pdf.

[78] Wagner, '§ 2 ProdHaftG para 15' (n 75); Oechsler, '§ 2 ProdHaftG para 64' (n 75).

[79] See, for example, Oechsler, '§ 2 ProdHaftG para 69' (n 75) (affirmative); and Wagner, '§ 2 ProdHaftG para 15' (n 75) (negative), both with further references.

[80] H Unberath, J Cziupka, and S Pabst, 'Art. 5 Rome II para 40' in T Rauscher (ed), *Europäisches Zivilprozess- und Kollisionsrecht (EuZPR/EuIPR), Kommentar, Volume 3: Rom I-VO, Rom II-VO* (4th ed. 2016); Remien, 'Article 5 Rome II para 4' (n 73).

[81] G Wagner, '§ 2 ProdHaftG para 21' in FJ Säcker and others (eds), *Münchener Kommentar zum Bürgerlichen Gesetzbuch, Volume 7: Schuldrecht – Besonderer Teil IV* (8th ed. 2020).

[82] Rome II, Article 14.

[83] Rome II, Article 5(2).

[84] See Sub-section II 3(b).

[85] Rome II, Articles 4(2) and 5(1).

Thirdly, if none of the above applies, Article 5(1) Rome II basically refers to the law of the state where the product was marketed, provided that the place of marketing coincides with one of three other territorial factors (the victim's habitual residence, the place where the product was acquired, the place of injury) and that the person claimed to be liable (usually the producer) could reasonably foresee the marketing of the product or a product of the same type in this country. Contrary to specific provisions on product liability, for example in Italy[86] or Switzerland,[87] Article 5(1) Rome II is not an alternative connection, but ranks the connecting factors in a hierarchical order. Firstly, the law applicable is that of the victim's habitual residence, provided that (1) it coincides with the place of marketing and (2) the producer does not succeed at proving that he could not foresee the marketing of this or a similar product in this country.[88] If one of those conditions (marketing, foreseeability) is not met, the law of the country in which the product was acquired applies, again subject to a coincidence with the place of marketing and the test of foreseeability.[89] If the applicable law cannot be determined at this stage, the law of the country in which the 'damage [read: injury] occurred', applies, if at least in this country the two additional requirements (marketing, foreseeability) are met.[90] If all of the three countries enumerated in Article 5(1) Rome II do not pass the test of foreseeability, the applicable law is that of the producer's habitual residence.

This rather unwieldy 'cascade system of connecting factors'[91] fails to achieve wholly convincing results. First, even after the Rome II Regulation has been in force now for more than a decade, it has not induced a single Member State, which is a party to the HCP, to denounce this convention. On the contrary, under Article 28 Rome II, the HCP takes precedence over the Rome II Regulation. The result is that, since 2009, Europeans have two different regimes on product liability conflicts which are both influenced by a similar methodology (grouping of contacts), but which do not yield uniform results in practice.

While Recital 20 explains that the 'conflict-of-law rule in matters of product liability should meet the objectives of fairly spreading the risks inherent in a modern high-technology society, protecting consumers' health, stimulating innovation, securing undistorted competition and facilitating trade,' it must be kept in mind that Article 5(1) Rome II is not limited to business-to-consumer (B2C) cases, but applies to business-to-business (B2B) cases as well.

Since the connecting factor that enjoys primacy in the basic rule[92] is relegated to the last rung of the ladder in cases of product liability,[93] drawing the line between general tortious liability and product liability is decisive in traffic accidents involving autonomous cars.[94] Thus, one may argue that there is a need for a special conflicts rule for those cases. A further complication arises from the above-mentioned fact that, in quite a number of member states, the law applicable to traffic accidents or product liability is still not determined by the Rome II Regulation, but by the pertinent Hague Conventions of the early 1970s (see Sub-section II.3(b)). Therefore, even an amendment to the Rome II Regulation would not create European legal unity in this regard.

[86] Article 63 of the Italian PIL Code (n 44).
[87] Article 135 of the Swiss PIL Code of 18 December 1987, SR 291, available at www.fedlex.admin.ch/eli/cc/1988/1776_1776_1776/de.
[88] Rome II, Article 5(1)(a).
[89] Rome II, Article 5(1)(b).
[90] Rome II, Article 5(1)(c).
[91] Rome II, Recital 20.
[92] Rome II, Article 4(1) place of damage.
[93] Rome II, Article 5(1)(c).
[94] See Kadner Graziano, 'Driverless cars' (n 6).

d. Special Rules in EU Law (Article 27 Rome II)

Pursuant to Article 27 Rome II, special EU conflicts rules take precedence over Rome II. In particular, the conflicts rules of the General Data Protection Regulation[95] may be relevant in cases involving AI.[96] In the course of the preparation of the Rome II Regulation, industry lobbies argued for codifying the 'country of origin'-approach as a choice-of-law rule.[97] While those attempts failed, Article 27 Rome II explicitly states that 'provisions of Community law which, in relation to particular matters, lay down conflict-of-law rules relating to non-contractual obligations' take precedence over the Regulation. Moreover, Recital 35 Rome II adds that the Regulation:

> should not prejudice the application of other instruments laying down provisions designed to contribute to the proper functioning of the internal market insofar as they cannot be applied in conjunction with the law designated by the rules of this Regulation. The application of provisions of the applicable law designated by the rules of this Regulation should not restrict the free movement of goods and services as regulated by Community instruments, such as ... [the] Directive on electronic commerce[98].

The precise reach of this exhortation is hard to define because the Directive on electronic commerce itself takes the somewhat schizophrenic position that it does not contain conflict-of-law rules,[99] while at the same time laying down the country-of-origin principle in its Article 3(1) and (2).[100] With regard to violations of rights of personality, a field not covered by Rome II, the CJEU tried to clarify matters as follows:[101]

> Article 3 of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ("Directive on electronic commerce"), must be interpreted as not requiring transposition in the form of a specific conflict-of-laws rule. Nevertheless, in relation to the coordinated field, Member States must ensure that, subject to the derogations authorized in accordance with the conditions set out in Article 3(4) of Directive 2000/31, the provider of an electronic commerce service is not made subject to stricter requirements than those provided for by the substantive law applicable in the Member State in which that service provider is established.

If the European legislature were to codify special conflicts rules on AI, such a regulation would not only supersede the Rome II Regulation pursuant to its Article 27, but arguably also take precedence over the Hague Conventions. The respective Articles 15 of the HCTA and the HCP state that the Hague Conventions shall not prevail over other Conventions 'in special fields' to

[95] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

[96] See JD Lüttringhaus, 'Das Internationale Datenprivatrecht: Baustein des Wirtschaftskollisionsrechts des 21. Jahrhunderts – Das IPR der Haftung für Verstöße gegen die EU-Datenschutzgrundverordnung' (2018) 117 *ZVglRWiss* 50.

[97] On the controversy, see von Hein, 'Abschluss' (n 3) 441; for a comprehensive theoretical and comparative analysis, see R Michaels, 'EU Law as Private International Law? Reconceptualising the Country-of-Origin Principle as Vested Rights Theory' (2006) 2 *J Priv Int'l L* 195.

[98] Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce), [2000] OJ L 178/1.

[99] E-Commerce Directive, Article 1(4).

[100] Cf. H Heiss and LD Loacker, 'Die Vergemeinschaftung des Kollisionsrechts der Außervertraglichen Schuldverhältnisse durch Rom II' (2007) 129 *Juristische Blätter* 613, 617, who criticise the Directive as 'wenig erhellend' ('little enlightening').

[101] CJEU, Joined Cases C-509/09 and C-161/10 *eDate Advertising GmbH and Others v X and Société MGN Limited* (25 October 2011).

which the contracting states are or may become parties. Although an EU Regulation is surely not a 'convention' within the technical meaning of those provisions, one may argue that Article 15 HCTA/HCP should apply by way of an analogy to any EU Regulation dealing with the law applicable to autonomous driving, for example.

4. *Contractual Obligations: The Rome I Regulation*

a. Scope

Complementing Rome II, the Rome I Regulation determines the law applicable to contractual obligations.[102] Mirroring the Rome II Regulation,[103] the notion of contractual obligation must be interpreted as an autonomous concept.[104] Thus, the Rome I Regulation designates the law applicable to so-called smart contracts, for example.[105] Likewise, the Rome I Regulation is of universal application as well.[106]

b. Choice of Law (Article 3 Rome I)

Party autonomy is largely permitted by Article 3 Rome I.[107] Consumers, however, must not be deprived of the protection accorded to them by the law of their habitual residence.[108]

c. Objective Rules (Articles 4 to 8 Rome I)

Usually, the habitual residence of the service provider determines the law applicable to contracts for services.[109] With regard to consumers, the law of the consumer's habitual residence applies under the conditions set out in Article 6(1) Rome I.[110]

d. Special Rules in EU Law (Article 23 Rome I)

Special conflicts rules in other EU legal instruments prevail over the Rome I Regulation.[111] There are occasional conflicts rules in older consumer directives;[112] however, the more recent directive on digital content and services does not contain any such rule.[113] On the contrary, Recital 80 of said directive explicitly states that '[n]othing in this Directive should prejudice the

[102] Rome I, Article 1(1).
[103] See Sub-section II 3(a).
[104] CJEU, Joined Cases C-359/14 and C-475/14 *'ERGO Insurance' SE, represented by 'ERGO Insurance' SE Lietuvos filialas, v 'If P&C Insurance' AS, represented by 'IF P&C Insurance' AS filialas (C-359/14), and 'Gjensidige Baltic' AAS, represented by 'Gjensidige Baltic' AAS Lietuvos filialas, v 'PZU Lietuva' UAB DK (C-475/14)* (21 January 2016), para 43.
[105] M Lehmann and F Krysa, 'Blockchain, Smart Contracts und Token aus der Sicht des (Internationalen) Privatrechts' [2019] *Bonner Rechtsjournal* 90; Wetenkamp, 'IPR und Digitalisierung' (n 6) 11; from a comparative point of view, see FA Schurr, 'Anbahnung, Abschluss und Durchführung von Smart Contracts im Rechtsvergleich' (2019) 118 *ZVglRWiss* 231.
[106] Rome I, Article 2.
[107] See M McParland, *The Rome I Regulation on the Law Applicable to Contractual Obligations* (2015) paras 9.01 *et seq* (herafter McParland, 'The Rome I Regulation').
[108] Rome I, Article 6(2); McParland, 'The Rome I Regulation' (n 107) paras 12.182–12.190.
[109] Rome I, Article 4(1)(b); Wetenkamp, 'IPR und Digitalisierung' (n 6) 20 *et seq*.
[110] McParland, 'The Rome I Regulation' (n 107) paras 12.01 *et seq*.
[111] Rome I, Article 23.
[112] See the enumeration in Article 46b(3) of the German EGBGB (n 36).
[113] Directive (EU) 2019/770 of the European Parliament and of the Council on certain aspects concerning contracts for the supply of digital content and digital services of 20 May 2019 [2019] OJ L 136/1.

application of the rules of private international law, in particular Regulations (EC) No 593/2008 and (EU) No 1215/2012 of the European Parliament and of the Council'.

III. THE DRAFT REGULATION OF THE EUROPEAN PARLIAMENT

1. *Territorial Scope*

With regard to substantive law, the draft regulation distinguishes between legally defined high-risk AI-systems[114] and other AI-systems involving a lower risk[115]. For high-risk AI-systems, the draft regulation would introduce an independent set of substantive rules providing for strict liability of the system's operator.[116] Further provisions deal with the amount of compensation,[117] the extent of compensation[118] and the limitation period.[119] The spatial scope of those autonomous rules on strict liability for high-risk AI-systems is determined by Article 2 DR, which reads as follows:

> 1. This Regulation applies on the territory of the Union where a physical or virtual activity, device or process driven by an AI-system has caused harm or damage to the life, health, physical integrity of a natural person, to the property of a natural or legal person or has caused significant immaterial harm resulting in a verifiable economic loss.
> 2. Any agreement between an operator of an AI-system and a natural or legal person who suffers harm or damage because of the AI-system, which circumvents or limits the rights and obligations set out in this Regulation, concluded before or after the harm or damage occurred, shall be deemed null and void as regards the rights and obligations laid down in this Regulation.
> 3. This Regulation is without prejudice to any additional liability claims resulting from contractual relationships, as well as from regulations on product liability, consumer protection, anti-discrimination, labour and environmental protection between the operator and the natural or legal person who suffered harm or damage because of the AI-system and that may be brought against the operator under Union or national law.

The unilateral conflicts rule found in Article 2(1) DR would prevail over the Rome II Regulation on the law applicable to non-contractual relations pursuant to Article 27 Rome II.[120] However, the Rome II Regulation still applies to additional liability claims mentioned in Article 2(3) DR. Moreover, Article 2(1) DR seems to limit the applicability of the draft regulation to cases where the harm was suffered on the territory of the European Union.[121] This stands in stark contrast with the principle of universal application that is one of the cornerstones of the Rome II Regulation.[122] If a high risk AI-system operated in Freiburg, Germany, for example, caused damage in Basel, Switzerland, the preconditions set out in Article 2(1) DR would not be met; thus, one would have to resort to the Rome II Regulation to determine the law applicable to the Swiss victim's claims.

[114] DR, Article 4.
[115] DR, Article 8.
[116] DR, Article 4.
[117] DR, Article 5.
[118] DR, Article 6.
[119] DR, Article 7.
[120] See Sub-section II 3(d).
[121] Pato (n 9) criticises the wording of Art. 2(1) DR as unclear and tends to favour an application of the DR 'where a court of a Member State is seized with a dispute involving damages caused by AI systems'.
[122] See Sub-section II 3(a).

2. *The Law Applicable to High Risk Systems*

Furthermore, it must be noted that Article 2(1) DR deviates considerably from the choice-of-law framework of Rome II. While Article 2(1) DR reflects the *lex loci damni* approach enshrined as the general conflicts rule in the Rome II Regulation,[123] one must not overlook the fact that product liability is subject to a special conflicts rule, namely Article 5 Rome II, which is considerably friendlier to the victim of a tort than the general conflicts rule.[124] This cascade of connections is evidently influenced by the desire to protect the mobile consumer from being confronted with a law that may be purely accidental from his point of view. The *lex loci damni*[125] may have neither a relationship with the legal environment that consumers are accustomed to[126] nor with the place where they decided to expose themselves to the danger possibly emanating from the product.[127] The rule reflects the presumption that a defective product will affect most consumers in the country where they are habitually resident. Insofar, Article 2(1) DR is, in comparison with the Rome II Regulation, friendlier to the *operator* of a high-risk AI-system than to the *consumer*.

Even if one limits the comparison between Article 2(1) DR and the Rome II Regulation to the latter's general rule,[128] it is striking that the DR does not adopt familiar approaches that allow for deviating from a strict adherence to *lex loci damni*. Contrary to Article 4(2) Rome II, where the person claimed to be liable and the person sustaining damage both have their habitual residence in the same country at the time when the damage occurs, Article 2 DR does not allow to apply the law of that country. Moreover, an escape clause such as Article 4(3) or Article 5(2) Rome II is missing in Article 2 DR. Finally, yet importantly, Article 2(2) DR bars any party autonomy with regard to strict liability for a high-risk AI-system, which deviates strongly from the liberal approach found in Article 14 Rome II.

3. *The Law Applicable to Other Systems*

Apart from the operator's strict liability for high-risk AI-systems, the draft regulation would introduce a fault-based liability rule for other AI-systems.[129] In principle, the spatial scope of the latter liability rule would also be determined by Article 2 DR as already described.[130] However, unlike the comprehensive set of rules on strict liability for high-risk systems, the draft regulation's model of fault-based liability is not completely autonomous. Rather, the latter type of liability contains important carve-outs regarding the amounts and the extent of compensation as well as the statute of limitations. Pursuant to Article 9 DR, those issues are left to the domestic laws of the Member States. More precisely, Article 9 DR states: 'Civil liability claims brought in accordance with Article 8(1) shall be subject, in relation to limitation periods as well as the amounts and the extent of compensation, to the laws of the Member State in which the harm or damage occurred.' Thus, we find a *lex loci damni* approach with regard to fault-based liability as well. Again, the principle of universal application[131] is discarded; contrary to the rules of Rome

[123] Rome II, Article 4.
[124] See Sub-section II 3(c).
[125] Rome II, Article 5(1)(c).
[126] His habitual residence: Rome II, Article 5(1)(a).
[127] Place of acquisition: Rome II, Article 5(1)(b).
[128] Rome II, Article 4.
[129] DR, Article 8.
[130] See Sub-section III 1.
[131] Rome II, Article 3.

II, Article 9 DR is a unilateral conflicts rule that only refers to 'the laws of the *Member State* in which the harm or damage occurred'. Moreover, all the modern approaches codified in the Rome II Regulation – the cascade of connecting factors for product liability claims, the common habitual residence rule, the escape clause, and party autonomy – are strikingly absent from Article 9 DR as well.

Finally, yet importantly, Article 9 DR leads to a split between the law applicable to the basis of liability, on the one hand, and the law applicable to limitation periods as well as the extent of compensation, on the other. This *dépeçage* stands in stark contrast with the general scope that Article 15 Rome II assigns to the *lex causae*. Pursuant to Article 15(a) Rome II, the law applicable to a non-contractual obligation under the Rome II Regulation covers both the basis and the extent of liability.[132] In addition, Article 15(h) Rome II provides that the law designated by the Rome II Regulation also applies to rules of prescription and limitation.[133] As *Axel Halfmeier* explains, 'the general tendency of the [Rome II] Regulation is to expand the reach of the *lex causae* and limit the role of the *lex fori* [because] the goal of the Rome Regulations is to produce harmony in results among the Member States' courts'[134] – the classic *Savignyan* goal of international decisional harmony mentioned above.[135] Of course, one has to take into account that Article 9 DR does not refer to the *lex fori*, but to the *lex loci damni*. Insofar, the rule does not offer any incentive for forum shopping. However, the unitary approach underlying Article 15 Rome II also serves the goal of 'avoiding the risk that the tort or delict is broken up in to several elements, each subject to a different law'.[136] Insofar, Article 15 Rome II aims at preventing the 'legal uncertainty' associated with applying different laws to a single case.[137] Particularly with regard to Article 15(h) Rome II, the Court of Justice of the EU (CJEU) 'pointed out that, in spite of the variety of national rules of prescription and limitation, Article 15(h) of the Rome II Regulation expressly makes such rules subject to the general rule on determining the law applicable'.[138] Creating a *dépeçage* between an autonomous rule on the conditions of liability, on the one hand, and the law applicable to the extent of damages and prescription issues, on the other, may lead to difficult questions of characterisation and adaptation. For example, the question may arise which particular rule of prescription of the *lex loci damni* shall apply if the latter law comprises various types of fault-based liability or calibrates the length of the prescription period depending on the degree of fault. In such a scenario, the court addressed would have to determine which domestic type of liability most closely corresponds to the model found in Article 8 DR – a task that may not be easy to fulfil. With regard to legal policy, it is hardly

[132] For a more detailed analysis, see I Bach, 'Article 15 Rome II para 1 et seq' in P Huber (ed), *Rome II Regulation – Pocket Commentary* (2011) (hereafter Bach, 'Article 15 Rome II para 1 et seq'); A Halfmeier, 'Article 15 Rome II paras 4–6' in GP Calliess and M Renner (eds), *Rome Regulations – Commentary* (3rd ed. 2020); G Palao Moreno, 'Article 15 Rome II paras 13–15' in U Magnus and P Mankowski (eds), *Rome II Regulation (European Commentaries on Private International Law)* (2019).

[133] For a closer analysis, see A Halfmeier, 'Article 15 Rome II paras 23–26' in GP Calliess and M Renner (eds), *Rome Regulations: Commentary* (3rd ed. 2020); G Palao Moreno, 'Article 15 Rome II para 23' in U Magnus and P Mankowski (eds), *Rome II Regulation (European Commentaries on Private International Law)* (2019).

[134] A Halfmeier, 'Article 15 Rome II para 2' in GP Calliess and M Renner (eds), *Rome Regulations: Commentary* (3rd ed. 2020); see also G Palao Moreno, 'Art. 15 Rome II para 2' in U Magnus and P Mankowski (eds) *Rome II Regulation (European Commentaries on Private International Law)* (2019) (prevention of forum shopping).

[135] See Sub-section II 1.

[136] CJEU, Case C-350/14 *Florin Lazar v Allianz SpA* (10 December 2015) para 29.

[137] Bach, 'Article 15 Rome II para 1 et seq.' (n 133).

[138] CJEU, Case C-149/18 *Agostinho da Silva Martins v Dekra Claims Services Portugal SA* (31 January 2019) para 33.

convincing to subject the issue of prescription to domestic laws because the periods codified in the Member States' laws have been criticised as being too short in light of the complexities of international cases.[139]

4. *Personal Scope*

The draft regulation, in principle, limits its personal scope to the liability of the operator alone.[140] Recital 9 of the resolution explains that the European Parliament

> [c]onsiders that the existing fault-based tort law of the Member States offers in most cases a sufficient level of protection for persons that suffer harm caused by an interfering third party like a hacker or for persons whose property is damaged by such a third party, as the interference regularly constitutes a fault-based action; notes that only for specific cases, including those where the third party is untraceable or impecunious, does the addition of liability rules to complement existing national tort law seem necessary.

Thus, for third parties, the conflicts rules of Rome II would continue to apply.

IV. EVALUATION

At first impression, it seems rather strange that a regulation on a very modern technology – AI – should deploy a conflicts approach that seems to have more in common with Joseph Beale's First Restatement of the 1930s[141] than with the modern and differentiated set of conflicts rules codified by the EU itself at the beginning of the twenty-first century (i.e. the Rome II Regulation). While the European Parliament's resolution, in its usual introductory part, diligently enumerates all EU regulations and directives dealing with substantive issues of liability, the Rome II Regulation is not mentioned once in the Recitals. One wonders whether the members of Parliament were aware of the European Union's *acquis* in the field of private international law at all.

V. SUMMARY AND OUTLOOK

In April 2020, the JURI Committee of the European Parliament presented a draft report with recommendations to the Commission on a civil liability regime for AI (see Sub-section I). The draft regulation proposed therein is noteworthy from a private international law perspective because it introduces new conflicts rules for AI. In this regard, the proposed regulation distinguishes between a rule delineating the spatial scope of its autonomous rules on strict liability for high-risk AI systems (Article 2 DR), on the one hand (see Sub-section III.2), and a rule on the law applicable to fault-based liability for low-risk systems (Article 9 DR), on the other hand (see Sub-section III.3.). The latter rule refers to the domestic laws of the Member State in which the harm or damage occurred. In this chapter, I have analysed and evaluated this proposal against the background of the already existing European regulatory framework on private international law, in particular the Rome II Regulation. In sum, compared with Rome II, the conflicts approach of the draft regulation would be a regrettable step backwards in many ways. On

[139] Kadner Graziano, 'Driverless cars' (n 6) 57.
[140] As legally defined in DR, Article 3(d)–(f).
[141] American Law Institute, *Restatement of the Law: Conflict of Laws* (1934).

21 April 2021, the European Commission presented its proposal for an 'Artificial Intelligence Act'.[142] However, this proposal contains neither rules on civil liability nor provisions on the pertinent choice-of-law issues. Thus, it remains to be seen how the relationship between the European Parliament's draft regulation and Rome II will be designed and fine-tuned in the further course of legislation.

[142] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021) 206 final.

PART IV

Fairness and Nondiscrimination in AI Systems

14

Differences That Make a Difference

*Computational Profiling and Fairness to Individuals**

Wilfried Hinsch

I. INTRODUCTION

The subject of this chapter is statistical discrimination by means of computational profiling. Profiling works on the basis of probability estimates about the future or past behavior of individuals who belong to a group characterized by a specific pattern of behavior. If statistically more women than men of a certain age abandon promising professional careers for family reasons, employers may expect women to resign from leadership positions early on and hesitate to offer further promotion or hire female candidates in the first place. This, however, would seem unfair to the well-qualified and ambitious young woman who never considered leaving a job to raise children or support a spouse. Be fair, she may urge a prospective employer, *Don't judge me by my group!*

Statistical discrimination is not new and not confined to computational profiling. Profiling, in all its variants – intuitive stereotyping, statistical in the old fashioned manner, or computational data mining and algorithm-based prediction – is a matter of information processing and a universal feature of human cognition and practice. It works on differences that make a difference. Profiling utilizes information about tangible features of groups of people, such as gender or age, to predict intangible (expected) features of individual conduct such as professional ambition. What has changed in the wake of technological progress and with the advent of Big Data and Artificial Intelligence (AI) is the effectiveness and scope of profiling techniques and with it the economic and political power of those who control and employ them. Increasing numbers of corporations and state agencies in some states are using computational profiling on a large scale, be it for private profit, to gain control over people, or other purposes.

Many believe that this development is not just a matter of beneficent technological progress.[1] Not all computational profiling applications promote human well-being, many undermine social justice. Profiling has become an issue of much public and scholarly concern. One major

* The phrase 'differences that make a difference' is taken from Gregory Bateson's *Steps towards an Ecology of Mind* (1972) where Bateson explains information in this way. I wish to express my gratitude for critical discussion and helpful commentary to Julian Sommerschuh, Silja Voeneky, and Gert Wagner.

[1] See among others BE Harcourt, *Against Prediction Profiling, Policing and Punishing in an Actuarial Age* (2007); AG Ferguson, *The Rise of Big Data Policing* (2017); V Eubanks, *Automating Equality: How High-Tech Tools Profile, Police and Punish the Poor* (2018); SU Noble, *Algorithms of Oppression. How Search Engines Reinforce Racism.* (2018); C O'Neill, *Weapons of Math Destruction: How Big Data Increases in Inequality and Threatens Democracy* (2017) S Zuboff, *The Age of Surveillance Capitalism* (2019), and KB Forrest, *When Machines Can Be Judge, Jury, and Executioner* (2021).

concern is police surveillance and oppression, another is the manipulation of citizens' political choices by means of computer programs that deliver selective and often inaccurate or incorrect information to voters and political activists. Yet another concern is the loss of personal privacy and the customization of individual life. The data mining and machine learning programs which companies such as Google, Facebook, and Amazon employ in setting up personal profiles run deep into the private lives of their users. This raises issues of data ownership and privacy protection. Profiles that directly target advertisements at receptive audiences thereby streamline and reinforce patterns of individual choice and consumption. This is not an outright evil and may not always be unwelcome. Nevertheless, it is a concern. Beliefs, attitudes, and preferences are increasingly shaped by computer programs which are operated and controlled in ways and by organizations that are largely, if not entirely, beyond our individual control.

The current agitation about 'algorithmic injustice' is fueled both by anxiety about, and fascination with, the remarkable development of information processing technologies that has taken place over the last decades. Against this backdrop of nervous attention is the fact that the ethical problems of computational profiling do not specifically relate to the computational or algorithmic aspect of profiling. They are problems of inappropriate discrimination based on statistical estimates in general. The main difference between discrimination based on biased computational profiling and discrimination based on false intuitive prejudice and stereotyping is scale and predictive power. The greater effectiveness and scope of computational profiling increases the impact of existing prejudices and, at many points, can be expected to deepen existing inequalities and reinforce already entrenched practices of discrimination. In a world in which playing on stereotypes and biases pays, both economically and politically, it is a formidable challenge to devise institutional procedures and policies for nondiscriminatory practices in the context of computational profiling.

This chapter is about unfair discrimination and the entrenchment of social inequality through computational profiling; it does not discuss concrete practical problems, however. Instead, it tackles a basic question of contemporary public ethics: what are the appropriate criteria of fairness and justice for assessing computational profiling appropriate for citizens who publicly recognize each other as persons with a right to equal respect and concern?[2]

Section I discusses the moral and legal concept of discrimination. It contains a critical review of familiar grounds of discrimination (*inter alia* ethnicity, gender, religion, and nationality) which figure prominently in both received understandings of discrimination and human rights jurisprudence. These grounds, it is argued, do not explain what and when discrimination is wrong (Section II 1 and 2). Moreover, focusing on specific personal characteristics considered the grounds of discrimination prevents an appropriate moral assessment of computational profiling. Section II, therefore, presents an alternative view which conceives of discrimination as a rule-guided social practice that imposes unreasonable burdens on persons (Sections II 3 and II 4). Section III applies this view to statistical and computational discrimination. Here, it is argued that statistical profiling is a universal feature of human cognition and behavior and not in itself wrongful discriminating (Section III 1).[3] Nevertheless, statistically sound profiles may prove objectionable, if not

[2] In this chapter, the terms 'fairness' and 'justice' will be used interchangeably for the most part. Depending on context, however, 'fairness' may, more specifically, refer to procedural features of profiling, 'justice' to substantive outcomes and empirical consequences. The phrase 'equal respect and concern' is taken from Ronald Dworkin's *Taking Rights Seriously* (1977).

[3] Unlike the German word '*Diskriminierung*' the English word 'discrimination' refers not exclusively to social conduct deemed morally objectionable. The term and its cognates are also used in a nonderogatory way. It is not necessarily a bad thing to have a discriminating mind or to make fine discriminations. 'Wrongful discrimination' or 'illicit discrimination' are not pleonasms. I shall use the phrases occasionally to highlight the moral disapproval of unfair discrimination.

inacceptable, for reasons of procedural fairness and substantive justice (Section III 2). It is argued, then, that the procedural fairness of profiling is a matter of degrees, and a proposal is put forth as regarding the general form of a fairness index for profiles (Section III 3).

Despite much dubious and often inacceptable profiling, the chapter concludes on a more positive note. We must not forget, for the time being, computational profiling is matter of conscious and explicit programming and, therefore, at least in principle, easier to monitor and control than human intuition and individual discretion. Due to its capacity to handle large numbers of heterogeneous variables and its ability to draw on and process huge data sets, computational profiling may prove to be a better safeguard of at least procedural fairness than noncomputational practices of disparate treatment.

II. DISCRIMINATION

1. *Suspect Grounds*

Discrimination is a matter of people being treated in an unacceptable manner for morally objectionable reasons. There are many ways in which this may happen. People may, for instance, receive bad treatment because others do not sympathize with them or hate them. An example is racial discrimination, a blatant injustice motivated by attitudes and preferences which are morally intolerable. Common human decency requires that all persons be treated with an equal measure of respect, which is incompatible with the derogatory views and malign attitudes that racists maintain toward those they hold in contempt. Racism is a pernicious and persistent evil, but it does not raise difficult questions in moral theory. Once it is accepted that the intrinsic worth of persons rests on human features and capacities that are not impaired by skin color or ethnic origin, not much reflection is needed to see that racist attitudes are immoral. Arguments to the contrary are based on avoidably false belief and unjustifiable conclusions.

However, some persons may still be treated worse than others in the absence of inimical or malign dispositions. Fathers, brothers, and husbands may be respectful of women and still deny them due equality in the contexts of household chores, education, employment, and politics. Discrimination caused by malign attitudes is a dismaying common phenomenon and difficult to eradicate. It is not the type of discrimination, however, that helps us to better understand the specific wrong involved in discrimination. Indeed, the very concept of statistical discrimination was introduced to account for discriminating patterns of social action that do not necessarily involve denigrating attitudes.[4]

Discrimination is a case of acting on personal differences that should not make a difference. It is a denial of equal treatment when, in the absence of countervailing reasons, equal treatment is required. The received understanding of discrimination is based on broadly shared egalitarian ethics. It can be summarized as follows: discrimination is adverse treatment that is degrading and violates a person's right to be treated with equal respect and concern. It is morally wrong because it imposes disparate burdens and disadvantages on persons who share characteristics like race, color, or sex, which on a basis of equal respect do not justify adverse treatment.

[4] See KJ Arrow, 'Models of Job Discrimination' in AH Pascal (ed), *Racial Discrimination in Economic Life* (1972) 83–102; KJ Arrow, 'What Has Economics to Say about Racial Discrimination?' (1998) 12 *Journal of Economic Perspectives* 91–100 and ES Phelps, 'The Statistical Theory of Racism and Sexism' (1972) 62 *American Economic Review* 659–661.

Discrimination is not unequal treatment of persons with these characteristics, it is unequal treatment because of them. The focus of the received understanding is on a rather limited number of personal attributes, for example, ethnicity, gender and sexual orientation, religious affiliation, nationality, disability, or age, which are considered to be the 'grounds of discrimination'. Hence, the question arises of which differences between people qualify as respectable reasons for unequal treatment, or rather, because there are so many valid reasons to make differences, which differences do not count as respectable reasons.

In a recruitment process, professional qualification is a respectable reason for unequal treatment, but gender, ethnicity, or national origin is not. In the context of policing people based on security concerns, the relevant difference must be criminal activity and not the ethnic or national origin of an alleged suspect. Admission to institutions of higher learning should be guided by scholarly aptitude and, again, not by ethnic or national origin, or any other of the suspect grounds of discrimination. The criteria which define widely accepted reasons for differential treatment (professional qualification, criminal activity, and scholarly ability) would seem to be contextual and depend on the specific purposes and settings. In contrast, the differences that should not make a difference like ethnicity or gender appear to be the same across a broad range of social situations.

In some settings and for some purposes, however, gender and ethnic or national origin could be respectable reasons for differential treatment, such as when choosing social workers or police officers for neighborhoods with a dominant ethnic group or immigrant population. Further, in the field of higher education, ethnicity and gender may be considered nondiscriminatory criteria for admission once it is taken into account that an important goal of universities and professional schools is to educate aspiring members of minority or disadvantaged groups to be future leaders and role models. Skin color may also be unsuspicious when choosing actors for screen plays, for example, casting a black actor for the role of Martin Luther King or a white actress to play Eleanor Roosevelt.[5] Nevertheless, selective choices guided by personal characteristics that are suspect grounds of discrimination appear permissible in specific contexts and in particular settings and seem impermissible everywhere else.

This is a suggestive take on wrongful discrimination which covers a broad range of widely shared intuitions about disparate treatment; however, it is misleading and inadequate as an account of discrimination. It is misleading in suggesting that the wrong of discrimination can be explained in terms of grounds of discrimination. It is inadequate in not providing operational criteria to draw a reasonably clear line between permissible and impermissible practices of adverse treatment. Not all selective actions based on personal characteristics that are considered suspect grounds of discrimination constitute wrongful conduct. It is impossible to decide whether a characteristic is a morally permissible reason for differential treatment without considering the purpose and context of selective decisions and practices. Therefore, a further criterion is needed to determine which grounds qualify as respectable reasons for differential treatment in specific settings and which do not.

[5] Skin color by itself may still seem incapable of justifying adverse treatment in conformity with a principle of equal respect and concern. This is true, however, for any other personal feature as well. It would be equally degrading for people with green eyes if they were treated worse than others based solely on eye color. No single feature or reason taken in isolation from other considerations justifies anything. All reasons for or against something are reasons only in the context of other reasons; an atomistic understanding of reasons must be avoided.

2. Human Rights

Reliance on suspect grounds for unequal treatment finds institutional support in international human rights documents. Article 2 of the 1949 Universal Declaration of Human Rights[6] contains a list of discredited reasons which became the template for similar lists in the evolving body of human rights law dealing with discrimination. It states: 'Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.'[7] Historically, the list makes sense. It reminds us of what, for a long period of time, was deemed acceptable for the denial of basic equal rights, and what must no longer be allowed to count against human equality. In terms of normative content, however, the list is remarkably redundant. If all humans are 'equal in dignity and rights' as the first Article of the Universal Declaration proclaims, all humans necessarily have equal moral standing and equal rights despite all the differences that exist between them, including, as a matter of course, differences of race, color, sex etc. Article 2 does not add anything to the proclamation of equal human rights in the Declaration. Further, the intended sphere of protection of the second Article does not extend beyond the sphere of the protection of the first. 'Discrimination' in the Declaration means denial of the equal rights promulgated by the Document.[8]

However, this is not all of it. Intolerable discrimination goes beyond treating others as morally inferior beings that do not have a claim to equal rights; and justice requires more than the recognition of equal moral and legal standing and a guarantee of equal basic rights. Article 26 of the International Covenant on Civil and Political Rights (ICCPR) introduces a more comprehensive understanding of discrimination. The first clause of the Article, however, contains the same redundancy found in the Universal Declaration. It states: 'All persons are equal before the law and are entitled without any discrimination to the equal protection of the law.' Equality before the law and the equal protection of the law are already protected by Articles 2, 16 and 17 of the ICCPR. Like all human rights, these rights are universal rights, and all individuals are entitled to them irrespective of the differences that exist between them. It goes, therefore, again without saying, that everyone is entitled to the protection of the law without discrimination.

It is the second clause of Article 26 which goes beyond what is already covered by the equal basic rights standard of the ICCPR: 'In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.' The broadening of scope in the quoted passage hinges upon an implicit distinction between equality before the law and equality through the law. In demanding 'effective protection against discrimination on any ground' without restriction or further qualification, Article 26 not only reaffirms the right to equality before the law (in its first clause), but also establishes a further right to substantive equality guaranteed through the law (in the second clause).

[6] United Nations, 'Universal Declaration of Human Rights' (12 October 1948) UN Doc. A/RES/3/217.

[7] The same list appears in Article 26 of the International Covenant on Civil and Political Rights (19 December 1966) 999 UNTS 171 and in Article 2 of the International Covenant on Economic, Social, and Cultural Rights (19 December 1966) 993 UNTS 3 (ICESCR), both of which became binding international law in 1976. An almost identical list can be found in Article 14 of the European Convention of Human Rights, Convention for the Protection of Human Rights and Fundamental Freedoms (4 November 1950) ETS No 005.

[8] See Article 7 which equates "equal protection of the law" with "equal protection against any discrimination in violation of this Declaration." Note also Article 6 of the international convention against racism from 1965 which defines racial discrimination as a violation of "human rights and fundamental freedoms contrary to this Convention."

Equality before the law is a matter of personal legal status and the procedural safeguards deriving from it. It is a demand of equal legal protection that applies (only) to the legal system of a society.[9] In contrast, equality through the law is the demand to legally ensure equality in all areas and transactions of social life and not solely in legal proceedings. Discrimination may then violate the human rights of a person in a two-fold manner. Firstly, it may be a denial of basic equal rights, including the right to equality before the law, promulgated by the Universal Declaration and the ICCPR. Alternatively, it may be a denial of due equality guaranteed by means of the law also beyond the sphere of equal basic rights and legal proceedings.[10]

If equality through the law goes further than what is necessary to secure equality before the law, a new complication for a human rights account of discrimination arises, not less disturbing than the charge of redundancy. Understood as equal legal protection of the basic rights of the Universal Declaration and the Covenant, non-discrimination simply means strict equality. All persons have the same basic rights, and all must be guaranteed the same legal protection of these rights. A strict equality standard of nondiscrimination, however, cannot plausibly be extended into all fields to be subjected to public authority and apply to social transactions in general. Not all unequal treatment even on grounds such as race, color, or sex is wrongful discrimination, and equating nondiscrimination with equal treatment *simpliciter* would tie up the human rights law of nondiscrimination with a rather radical and indefensible type of legalistic egalitarianism.

Elucidation concerning the equal treatment requirement of nondiscrimination can be found in both the 1965 Convention against Racial Discrimination (ICERD)[11] and in the 1979 Convention against gender discrimination (CEDAW).[12] The ICERD defines discrimination as 'any distinction, exclusion or preference' with the purpose or effect of '... nullifying or impairing the recognition, enjoyment or exercise, on an equal footing of human rights and fundamental freedoms' (Article 1). The CEDAW refers to 'a basis of equality' between men and women (Article 1) and demands legislation that ensures "... full development and advancement of women for the purpose of guaranteeing them the exercise and enjoyment of human rights and fundamental freedoms on a basis of equality with men" (Article 3). Following these explications, differential treatment even on one of the grounds enumerated in the ICCPR would not *per se* constitute illicit discrimination. It would only do so if it proved incompatible with the exercise and enjoyment of basic human rights on an equal basis or equal footing.

Both 'equal basis' and 'equal footing' suggest an understanding of nondiscrimination that builds on a distinction between treating people as equals, in other words, with equal respect and concern, but still differently and treating people equally. As a matter of equal basic rights protection, nondiscrimination means strict equality, literally speaking, equal treatment. As a matter of protection against disparate treatment that does not violate people's basic rights, nondiscrimination would still require that everyone is treated as equal but not necessarily treated

[9] Still, equality before the law is not merely formal: Substantive legal regulation of judicial proceedings is needed to ensure equality before the law and equal legal protection. After all, the legal system of a society is itself a field of social transactions.

[10] This interpretation aligns with the Human Rights Committee's understanding of Article 26 to which *Sarah Joseph* and *Melissa Castan* refer in their commentary on the Covenant. "In the view of the Committee, Article 26 does not merely duplicate the guarantee already provided for in Article 2 but provides an autonomous right. It prohibits discrimination in law or in fact any field." (See S Joseph and M Castan (eds), *The International Covenant on Civil and Political Rights: Cases, Materials, and Commentary* (2013) section 23.15 (hereafter Joseph and Castan, *The ICCPR*)). If Article 26 prohibits discrimination not only 'in law' but in 'any field' under supervision and protection of public authorities, it effectively prohibits adversely unequal treatment beyond the denial of equality before the law and, quite generally, the denial of equal basic rights.

[11] International Convention on the Elimination of All Forms of Racial Discrimination (7 March 1966) 660 UNTS 1.

[12] Convention on the Elimination of All Forms of Discrimination against Women (18 December 1979) 1249 UNTS 1.

equally. Given adequate legal protection against the violation of basic human rights, not all adversely unequal treatment would then constitute a violation of the injunction against discrimination of Article 26 of the ICCPR.

The distinction between equal treatment and treatment as equals provides a suitable framework of moral reasoning and public debate and perhaps also a suggestive starting point of legal argument. 'Equal respect' and 'equal concern' effectively capture a broadly shared intuitive idea of what it takes to enjoy basic rights and liberties on the 'basis of equality'. Yet, these fundamental distinctions and ideas allow for differing specifications. On their own and without further elaboration, they do not provide a reliable basis for the consistent and predictable right to nondiscrimination. The formula of equal respect and concern is a matter of contrary interpretations in moral philosophy. Some of these interpretations are of a classical liberal type and ultimately confine the reach of antidiscrimination norms to the sphere of elementary basic rights protection. Other interpretations, say of a utilitarian or *Rawlsian* type, and extend the demand of protection against supposed discriminatory decisions and practices beyond elementary basic rights protection.

The problem here is not the absence of an uncontested moral theory specifying terms of equal respect and concern. While moral philosophy and public ethics have long been controversial, legal and political theory found ways to accommodate not only religious but also moral pluralism. The problem is that a viable human rights account of discrimination must draw a reasonably clear line between permissible and impermissible conduct, and this presupposes a rather specific understanding of what it means to treat people on the 'basis of equality' or with equal respect and concern. The need to specify the criteria of illicit discrimination with recourse to a requirement for equal treatment based on human rights thus leads right into the contested territory of moral philosophy and competing theories of justice. Without an involvement in moral theory, a human rights account would seem to yield no right to nondiscrimination which is reasonably specific and nonredundant, given reasonable disagreement in moral theory, it seems impossible to specify such a right in a way that could not be reasonably contested.

The ambiguities of a human right to nondiscrimination also becomes apparent elsewhere. While not all differential treatment on the grounds of race, color, sex etc. is wrongful discrimination, not all wrongful discrimination is discrimination on these grounds. Article 26 prohibits discrimination not only when it is based on one of the explicitly mentioned attributes but on any ground such as race, color, sex etc. or other status. Therefore, the question arises of how to identify the grounds of wrongful discrimination and of what qualifies as an 'other status.' The ICCPR does not answer the question and the UN Human Rights Committee seems to be at a loss when it comes to deciding about 'other grounds' and 'other status' in a principled manner.[13]

[13] See again two references in *Joseph* and *Castan's* commentary: "The HRC may view certain grounds of distinction as inherently more suspect and deserving of greater scrutiny than other grounds. [...] It seems intrinsically more important to guard against discrimination on the grounds such as [...] nationality, sexuality, age, or disability, than it is to protect against discrimination on other grounds" (Joseph and Castan, *The ICCPR* (n 11) section 23.36). 'The HRC has not issued a detailed consensus on the meaning of "any other status," preferring to decide on a case-by-case basis whether a complaint raises a relevant ground of discrimination' (Joseph and Castan, *The ICCPR* (n 11) section 23.27). Are these quotes concessions of juridical defeat? The grounds of discrimination in Article 26 clearly deserve attention. And clearly, other grounds not mentioned in the article but now generally accepted as reasons of wrongful discrimination like physical impairment, sexual orientation, or age must also be critically attended to. We must not conclude from this, however, that suspect grounds are 'inherently' more suspect than others or that it is 'intrinsically' more important to guard against them. Racism is not intrinsically related to race and sexism not to sex. Race, color, or sex do not attract by themselves unfair treatment. Much illicit discrimination proceeds along the lines of the personal characteristics mentioned in Article 26. However, this is due to contingent social, cultural, political, economic, or other causes and not to an 'intrinsic' quality of these characteristics.

Ethnicity and gender, for instance, figure prominently in inacceptable practices of disparate treatment. However, these practices are not inacceptable because they are guided by considerations of ethnicity or gender. Racial or gender discrimination are not paradigm cases of illicit discrimination because ethnicity and gender could never be respectable reasons to treat people differently or impose unequal burdens. They are paradigm cases because ethnic and gender differences, as a matter of historical fact, inform social practices that are morally inacceptable. What then makes a practice of adversely unequal treatment that is guided by ethnicity or gender or, indeed, any other personal characteristic morally inacceptable?

In their commentary about the ICCPR, *Joseph* and *Castan* are candid about the difficulty of ascribing 'common characteristics' for the 'grounds' in Article 26.[14] It is always difficult if not impossible to add tokens to a list of samples in a rule-guided way without making contestable assumptions. Still, we normally have some indication from the enumerated samples. In the case of … table, chair, cupboard …, for instance, we have a conspicuous classificatory term, 'furniture', as a common denominator that suggests proceeding with 'couch' or 'floor lamp' but not with 'seagull'. What would be the common denominator of … race, color, sex … except that these personal attributes are grounds of wrongful discrimination?[15] If exemplary historical cases are meant to guide the identification of suspect grounds, however, these grounds are no longer independent criteria that explain why these cases provide paradigm examples of discrimination and we may wonder which other types of disparate social treatment may be considered wrongful discrimination as well.

Contrary to appearance, Article 26 provides no clue as regarding the criteria of wrongful discrimination. Not all adverse treatment on the grounds mentioned in the article is wrongful discrimination and not all wrongful discrimination is discrimination on these grounds. Race, color, sex, etc. have been and continue to be grounds of intolerable discrimination. Adverse treatment based on these characteristics, therefore, warrants suspicion and vigilance.[16] However, since it is a matter of purpose and context whether adverse treatment based on a personal feature is compatible with equal respect and concern, we still need an account of the conditions under which it constitutes wrongful discrimination. Moreover, an explanation why adverse treatment is wrong under these conditions is also required. Suspected grounds of discrimination and the principle of equal basic rights offer neither.[17]

3. *Social Identity or Social Practice*

Discrimination has many faces. It may be personal – one person denying equality to another – or impersonal where it is a matter of biased institutional measures and procedures. It may also be direct or indirect, intended or unintended. But it is never a matter of isolated individual

[14] Joseph and Castan, *The ICCPR* (n 11) section 23.36.

[15] See again Joseph and Castan: "Perhaps the most common characteristic of an important 'ground' is that the 'ground' describes a group which has historically suffered from unjustifiable discrimination and is therefore especially vulnerable to such treatment." See Joseph and Castan, *The ICCPR* (n 11) section 23.36.

[16] See *Antje von Ungern-Sternberg's* discussion of the suspect grounds of discrimination in A von Ungern-Sternberg, 'Religious Profiling, Statistical Discrimination and the Fight against Terrorism' in R Uerpmann-Wittzack (ed.), *Religion and International Law* (2017) 191–211.

[17] The policy of the Human Rights Committee, reported by *Joseph* and *Castan*, to decide in a case-by-case manner on the 'grounds of discrimination' and on 'other status' may not yield unreasonable decisions in specific cases. Nevertheless, it raises vexing questions: how does the committee decide without explicit criteria whether a personal characteristic, which in a given context functions as a reason for adversely differential treatment, is a ground of illicit discrimination? Or how does it ensure the consistency of its case-by-case decisions over time; and how does it respond to charges of ill-conceived discrimination?

wrongdoing. Discrimination is essentially social. It occurs when members of one group, directly or indirectly, intentionally or unintentionally, consistently treat members of another group badly, because they perceive them as deficient in some regard. Discrimination requires a suitable context and takes place against a backdrop of socially shared evaluations, attitudes, and practices. To emphasize the social nature of discrimination not only reflects linguistic usage, it also helps us to understand what is wrong with it and to shift the attention from lists of suspect grounds to the practices and burdens of discrimination.

An employer who does not hire a well-qualified applicant because she is a woman may be doing something morally objectionable for various reasons: a lack of respect, for instance, or prejudice. If he or she were the only employer in town, however, who refused to hire women, or one of only a few, their hiring decision, I suggest, though morally objectionable, would not consitute illicit discrimination. In the absence of other employers with similar attitudes and practices, their bias in favor of male workers, though objectionable and frustrating for female candidates, does not lead to the special kind of burdens and disadvantages that characterize discrimination. Indeed, a dubious gender bias of only a handful of people may not result in serious burdens for women at all. Rejected candidates would easily find other jobs and work somewhere else. It is only the cumulative social consequences of a prevailing practice of gender-biased hiring that create the specific individual burdens of discrimination. There is a big difference between being rejected for dubious reasons at some places and being rejected all over the place.

Consider, in contrast, individual acts of wrongdoing which are not essentially social because they do not depend on the existence of practices that produce cumulative outcomes which disparately affect others. We may maintain, for instance (pace *Kant*) that false promising is only wrong if there is a general practice of promise-keeping. However, it would seem odd to claim that an individual act of false promising is only wrong if there is a general practice of promise-breaking with inacceptable cumulative consequences for the involved people. Unlike acts of wrongful promise-breaking, acts of wrongful discrimination do not only depend upon the existence of social practices – this may be true for promise breaking as well. They crucially hinge upon the existence of practices with cumulative consequences which impose burdens on individuals that only exist because of the practice. This suggests a social practice view of discrimination.

Social practices are regular forms of interpersonal transactions based on rules which are widely recognized as standards of appropriate conduct among those who participate in the practice. They rest on publicly shared beliefs and attitudes. The rules of a practice define spheres of optional and nonoptional action and specify types of advisable as well as obligatory conduct. They also define complementary positions for individuals with different roles who participate in the practice or who are subjected to it or indirectly affected by it. Practices may or may not have a commonly shared purpose, but they always have cumulative and noncumulative consequences for the persons involved, and any plausible moral assessment must, in one way or another, take these consequences into account.

Social practices of potentially wrongful discrimination are defined by the criteria which guide the discriminating choices of the participants, in other words, the specific generic personal characteristics which (a) function as the grounds of discrimination and (b) identify the group of people who are targeted for adverse treatment. This gives generic features of persons a central place in any conception of discrimination. These characteristics are not 'grounds of discrimination', however, because they adequately explain the difference between differential treatment that is morally or legally unobjectionable and treatment that constitutes discrimination. We have

seen that by themselves, they do not provide suitable criteria for the moral appraisal of disparate treatment. Instead, they identify the empirical object of moral scrutiny and appraisal, in other words, social practices of differential treatment that impose specific burdens and disadvantages on the group of persons with the respective characteristics.

To reiterate, the wrong of discrimination is not a wrong of isolated individual conduct. It only takes place against the backdrop of prevailing social practices and their cumulative consequences and, for this reason, it cannot be fully explained as a violation of principles of transactional or commutative justice. This leads into the field of distributive justice. Principles of transactional justice, like the moral prohibition of false promising or the legal principle *pacta sunt servanda*, presuppose individual agency and responsibility. They do not apply to uncoordinated social activities or cumulative consequences of individual actions that transcend the range of individual control and foresight. Clearly, social practices of discrimination only exist because there are individual agents who make morally objectionable discriminating choices. The choices they make, however, would not be objectionable if not for the cumulative consequences of the practice of which they are a part and to which they contribute. We therefore need standards for the assessment of the cumulative distributive outcomes of individual action, in other words, standards of distributive justice that do not presuppose individual wrongdoing but rather explain it. We come back to this in the next section.

There is another train of thought which also explains the essentially social character of discrimination though, not in terms of shared practices of adverse treatment but in terms of disadvantaged social groups. Most visibly discrimination is directed against minority groups and the worse-off members of society. This may well be seen to be the reason why discrimination is wrong.[18] Is it, then, a defining feature of discrimination that it targets specific types of social groups? The list of the suspect grounds of discrimination in article 26 of the ICCPR may suggest that it is because the mentioned ground appear to identify groups that fit this description.

Thomas Scanlon and *Kasper Lippert-Rasmussen* have followed this train of thought in slightly different ways. By *Scanlon's* account, discrimination disadvantages 'members of a group that has been subject to widespread denigration and exclusion'.[19] On *Lippert-Rasmussen's* account, discrimination is denial of equal treatment for members of 'socially salient groups' where a group is socially salient if 'membership of it is important to the structure of social transactions across a wide range of social contexts'.[20] Examples of salient groups are groups defined by personal characteristics like sex, race, or religion, characteristics which, unlike having green eyes, for instance, make a difference in many transactions and inform illicit practices in various settings; salient groups, for this reason, inform social identities. This accords well with common understandings of discrimination and explains the social urgency of the issue: it is not only individuals being treated unfairly for random reasons in particular circumstances, it is groups of people who are regularly treated in morally objectionable ways across a broad range of social transactions and for reasons that closely connect with their personal identity and self-perception.

Still, neither the intuitive notion of denigrated and excluded groups nor the more abstract conception of salient groups adequately explain what is wrong with discrimination. Both approaches run the risk of explaining discrimination in terms of maltreatment of discriminated

[18] The notion of a minority, however, though of great political importance, is rather an obstacle to an adequate understanding of discrimination. Women are not a minority and still subjected to unfair discrimination. With immigrants, all depends on the numbers, and we must not forget the discrimination of majorities in the wake of imperialism and colonial rule.

[19] T Scanlon, *Moral Dimensions* (2008) 74.

[20] K Lippert-Rasmussen, 'The Badness of Discrimination' (2006) 9 *Ethical Theory and Moral Practice* 167, 168 *et seq*.

groups. More importantly, both lead to a distorted picture of the social dynamics of discrimination. While the most egregious forms of disparate treatment track personal characteristics that do define excluded, disadvantaged, and 'salient groups', it is not a necessary feature of discrimination that it targets only persons who belong to and identify with groups of this type.

Following *Scanlon* and *Lippert-Rasmussen*, discrimination presupposes the existence of individuals who are already (unfairly, we assume) disadvantaged in a broad range of social transactions. Discrimination becomes a matter of piling up unfair disadvantages – a case of adding insult to injury one may say. This understanding, however, renders it impossible to account for discriminating practices that lead to exclusion, disadvantage, and denigration in the first place. *Lippert-Rasmussen's* understanding of salient groups creates a blind spot for otherwise well-researched phenomena of context-specific and partial forms of discrimination which do not affect a broad range of a person's social transactions and still seriously harm them in a particular area of life. Common sense suggests and social science confirms that discrimination may be contextual, piecemeal, and, in any case, presupposes neither exclusion or disadvantage nor prior denigration of groups of people.[21]

It is an advantage of the practice view of discrimination that it is not predicated on the existence of disadvantaged social groups. Based on the practice view, the elementary form of discrimination is neither discrimination of specific types of social groups that are flagged in one way or another as excluded or disadvantaged, nor is it discrimination because of group membership or social identity. It is discrimination of individual persons because of certain generic characteristics – the grounds of discrimination – that are attributed to them. Individuals who are subjected to discriminating practices due to features which they share with other persons are, because of this, also members of the group of people with these features. However, group membership here means nothing more than to be an element of a semantic reference class, that is, the class of individuals who share a common characteristic. No group membership in any sociologically relevant sense or in *Lippert-Rasmussen's* sense is implied; nor is there any sense of social identity or prior denigration and exclusion.[22]

To appreciate the relevance of group membership and social identities in the sociological sense of these words, we need to distinguish between what constitutes the wrong of illicit discrimination in the first place and what makes social practices of discrimination more or less harmful under some conditions than under others. Feelings of belonging to a group of people with a shared sense of identity who have been subjected to unfair discrimination for a long time and who are still denied due equality intensifies the individual burdens and harmful effects of discrimination. It heightens a person's sense of being a victim not of an individual act of wrongdoing but of a long lasting and general social practice. Becoming aware that one is subjected to adverse treatment because of a feature that one shares with others, in other words, becoming aware that one is an element of the reference class of the respective feature, also means becoming aware of a 'shared fate', the fate of being subjected to the same kind of disadvantages for the same kind of reasons. And this in turn will foster sympathetic identification

[21] For the empirical findings of experimental social psychology concerning context-specific and partial forms of discrimination that do not track social identities see J Holroyd, 'The Social Psychology of Discrimination' in K Lippert-Rasmussen (ed), *The Routledge Handbook of the Ethics of Discrimination* (2018) 381–393.

[22] The persons who are subjected to discriminating practices in this elementary sense do not even have to be aware that there are others who are discriminated against because of the same characteristics, and they may have never been the victims of illicit discrimination before. This is a point of some importance when assessing the moral permissibility of computational profiling that targets highly specific groups of individuals who are identified by a great number of non-salient characteristics and who may not even know that they have these characteristics or that they share them with others.

with other group members and perhaps also feelings of belonging. Moreover, it creates a shared interest, viz. the interest not to be subjected to adversely discriminatory practices, which in turn may contribute to the emergence of new political actors and movements.

4. *Disparate Burdens*

It is hard to see that we should abstain from discriminatory conduct if it did not cause harm. In all social transactions, we continuously and inevitably spread uneven benefits and burdens on others by exercising preferential choices. Much of what we do to others, though, is negligible and cannot be reasonably subjected to moral appraisal or regulation; and much of what we do, though not negligible, is warranted by prior agreements or considerations of mutual benefit. Finally, much adversely selective behavior does not follow discernible rules of discrimination and may roughly be expected to affect everybody equally from time to time. Wrongful discrimination is different as it imposes in predictable ways, without prior consent or an expectation of mutual benefit, burdens and disadvantages on persons which are harmful.[23]

Not all wrongful harming is discrimination. Persons discriminated against are not just treated badly, they are treated worse than others. A teacher who treats all pupils in his class with equal contempt behaves in a morally reprehensible way, but he cannot be charged on the grounds of discrimination. The harm of discrimination presupposes an interpersonal disadvantage or comparative burden, not just additional burdens or disadvantages. It is one thing to be, like all others, subjected to inconvenient security checks at airports and other places, it is another thing to be checked more frequently and in more disagreeable ways than others. To justify a complaint of wrongful treatment, the burdens of discrimination must also be comparative and interpersonal in a further way. Adverse treatment is not generally impermissible if it has a legitimate purpose. It is only wrongful discrimination if it imposes unreasonable burdens and disadvantages on persons, burdens and disadvantages that cannot be justified by benefits that otherwise derive from it.

We thus arrive at the following explanation of wrongfully discriminating practices in terms of unreasonable or disproportionate burdens and disadvantages: a social practice of adverse treatment constitutes wrongful discrimination if following the rules of the practice – acting on the 'grounds' of discrimination – imposes unreasonable burdens on persons who are subjected to it. Burdens and disadvantages of a discriminatory practice are unreasonable or disproportionate if they cannot be justified by benefits that otherwise accrue from the practice on a basis of equal respect and concern which gives at least equal weight to the interests of those who are made worse off because of the practice.

It is an advantage of the unreasonable burden criterion that we do not have to decide whether the wrong of discrimination derives from the harm element of adverse discrimination or from the fact that the burdens of discrimination cannot be justified in conformity with a principle of equal respect and concern. Both the differential burden and the lack of a proper justification are necessary conditions of discrimination. Hence, there is no need to decide between a harm-based and a respect-based account of discrimination. If it is agreed that moral justifications must proceed on an equal respect basis, all plausible accounts of discrimination must seem to

[23] Naturally, people have different ideas about nonnegligible burdens. There are limits, though, as to what may count as negligible among humans, given the fragility and vulnerability of our existence. Still, there is no hard and fast line between negligible burdens or disadvantages and serious harm. Complaints about discrimination are, therefore, bound to be controversial in many cases. In any case, a principle of nondiscrimination presupposes a commonly recognized threshold of inacceptable burdens if it is meant to provide a viable standard of public ethics.

combine both elements. There is no illicit discrimination if we either have an unjustified but not serious burden or a serious yet justified burden.

The criterion of unreasonable burdens may appear to imply a utilitarian conception of discrimination,[24] and, indeed, combined with this criterion, the practice view yields a consequentialist conception of discrimination. However, this conception can be worked out in different ways. The goal must not be to maximize aggregate utility and balancing the benefits and burdens of practices does not need to take the form of a cost-benefit-analysis along utilitarian lines. The idea of an unreasonable burden can also be spelled out – and more compellingly perhaps – along Prioritarian or *Rawlsian* lines, giving more weight to the interests of disadvantaged groups.[25] We do not need to take a stand on the issue, however, to explain the peculiar wrong of discrimination. On the proposed view, it consists in an inappropriate social distribution of benefits and burdens. It is a wrong of distributive justice.

One may hesitate to accept this view. It seems to omit what makes discrimination unique, and to explain why people often feel more strongly about discrimination than about other forms of distributive injustice. What is special about discrimination, however, is not an entirely new kind of wrong; instead, it is the manner in which a distributive injustice comes about, the way in which an unreasonable personal burden is inflicted on a person in the pursuit of a particular social practice. Not all distributive injustice is the result of wrongful discrimination, but only injustice that occurs as the predictable result of an on-going practice which is regulated by rules that track personal characteristics which function as grounds of discriminating choices.

Consider, by way of contrast, the gender pay gap with income inequality in general. In a modern economy, the primary distribution of market incomes is the cumulative and unintended result of innumerable economic transactions. Even if all transactions conformed to principles of commutative or transactional justice and would be unassailable in terms of individual intentions, consequences, and responsibilities, the cumulative outcomes of unregulated market transactions can be expected to be morally inacceptable. Unfettered markets tend to produce fabulous riches for some people and bring poverty and destitution to many others. Still, in a complex market economy, it will normally be impossible to explain an unjust income distribution in terms of any single pattern of transactions or rule-guided practice. To address the injustice of market incomes we, therefore, need principles of a specifically social, or distributive justice, which like the *Rawlsian* Difference Principle[26] apply to overall statistical patterns of income (or wealth) distribution and not to individual transactions.

Consider now, by way of contrast, the inequality of the average income of men and women. The pay gap is not simply the upshot of a cumulative but uncoordinated – though still unjust – market process. Our best explanation for it is gender discrimination, the existence of rule-guided social practices, which consistently in a broad range of transactions put women at an unfair disadvantage. Even though the gender pay gap, just like excessive income inequality in general, is a wrong of distributive injustice, it is different in being the result of a particular set of social practices that readily explain its existence.

[24] At any rate, it is not apparent that the proposed view is incompatible with utilitarianism. Much depends on whether utilitarian principles are consistent with the more general principle of equal respect and concern.

[25] Note the difference between (a) defining discrimination as adverse treatment of disadvantaged groups and (b) claiming, as a matter of substantive moral argument, that interests of disadvantaged parties should be given extra weight in assessing social practices of discrimination.

[26] The Difference Principle requires that the overall distribution of income and wealth in a society maximize the lifetime prospects of the worst-off.

This account of wrongful disparate treatment, however, does not accord well with a human rights theory of discrimination. A viable human right to nondiscrimination presupposes an agreed upon threshold notion of nonnegligible burdens and a settled understanding of how to balance the benefits and burdens of discriminatory practices in appropriate ways. If the interpersonal balancing of benefits and burden, however, is a contested issue and subject to reasonable disagreement in moral philosophy, that which is protected by a human right to nondiscrimination would also seem to be a subject of reasonable disagreement. Given the limits of judicial authority in a pluralistic democracy, and given the need of democratic legitimization for legal regulations that allow for reasonable disagreement, this suggests that antidiscrimination rights should not be seen as prelegislative human rights but more appropriately as indispensable legal elements of a just social policy the basic terms of which are settled by democratic legislation and not by the courts.

This is not to deny that there are human rights – the right to life, liberty, security of the person, equality before the law – the normative core of which can be determined in ways that are arguably beyond reasonable dissent. For these rights, but only for them, it may be claimed that their violation imposes unreasonable burdens and, hence, constitutes illicit discrimination without getting involved with controversial moral theory. For these rights, however, a special basic right of nondiscrimination is superfluous, as we have seen in Section I.2. (If all humans have the same basic rights, they have these rights irrespective of all differences between them and it goes without saying that these rights have to be equally protected ['without any discrimination'] for all of them.) And once we move beyond the equal basic rights into the broader field of protection against unfair social discrimination in general, the determination of unreasonable burdens is no longer safe from reasonable disagreements. Institutions and officials in charge of enforcing the human right of nondiscrimination would then have a choice, which, among the reasonable theories, would be used to assess the burdens of discrimination. Clearly, they must be expected to come up with different answers. Quite independent from general concerns about the limits of judicial discretion and authority, this does not accord well with an understanding of basic rights as moral and legal standards which publicly establish a reasonably clear line between what is permissible and what is impermissible and conformity which can be consistently enforced in a reasonably uniform way over time.[27]

Let us briefly summarize the results of our discussion so far: firstly, a social practice of illicit discrimination is defined by rules that trace personal characteristics, the grounds of discrimination, which function as criteria of adverse selection.

Secondly, the cumulative outcome of on-going practices of discrimination leads to unequal burdens and disadvantages which adversely affect persons who share the personal characteristics specified by the rules of the practice.

Thirdly, the nature and weight of the burdens of discrimination are largely determined by the cumulative effects which an on-going practice of discrimination produces under specific empirical circumstances.

[27] Note that this line of argument does not presuppose that we are able to clearly distinguish between types of (human rights) protection against discrimination that is subject to reasonable disagreement and others that are not. Wherever the line between the core of basic human rights protection and the broader field of protection against objectionable disparate treatment is drawn, core protection implies a standard of strict equality which cannot be defended for the broader field. We need a more inclusive understanding of equality, something like 'on a basis of equality,' or 'on an equal footing,' which invariably will be subject to contrary reasonable interpretations.

Fourthly, discrimination is morally objectionable or impermissible, if discriminating in accordance with its defining set of criteria imposes unreasonable burdens on persons who are adversely subjected to it.

III. PROFILING

1. *Statistical Discrimination*

Computational profiling based on data mining and machine learning is a special case of 'statistical discrimination'. It is a matter of statistical information leading to, or being used for, adverse selective choices that raise questions of fairness and due equality. Statistics can be of relevance for questions of social justice and discrimination in various ways. A statistical distribution of annual income, for instance, may be seen as a representation of injustice when 10% of the top earners receive 50% of the national income while the bottom 50% receive only 10%. Statistics can also provide evidence of injustice, for example, when the numerical underrepresentation of women in leadership positions indicates the existence of unfair recruitment practices. And, finally, statistical patterns may (indirectly) be causes of unfair discrimination or deepen inequalities that arise from discriminatory practices. If more women on average than men drop out of professional careers at a certain age, employers may hesitate to promote women or to hire female candidates for advanced management jobs. And, if it is generally known that statistically, for this reason, few women reach the top, girls may become less motivated than boys to acquire the skills and capacities necessary for top positions and indirectly reinforce gender stereotypes and discrimination.

Much unfair discrimination is statistical in nature not only in the technical or algorithmic sense of the word. It is based on beliefs about personal dispositions and behavior that allegedly occur frequently in groups of people who share certain characteristics such as ethnicity or gender. The respective dispositional and behavioral traits are considered typical for members of these groups. Negative evaluative attitudes toward group members are deemed justified if the generic characteristics that define group membership correlate with unwanted traits even when it is admitted that, strictly speaking, not all group members share them.

Ordinary statistical discrimination often rests on avoidable false beliefs about the relative frequency of unwanted dispositional traits in various social groups that are defined by the characteristics on the familiar lists of suspected grounds '… race, color, sex …' Statistical discrimination need not be based, though, on prejudice and bias or false beliefs and miscalculations. Discrimination that is statistical in nature is a basic element of all rational cognition and evaluation; statistical discrimination in the technical sense with organized data collection and algorithmic calculations is just a special case. Employers may or may not care much about ethnicity or gender, but they have a legitimate interest to know more about the future contribution of job candidates to the success of their business. To the extent that tangible characteristics provide sound statistical support for probability estimates about the intangible future economic productivity of candidates, the former may reasonably be expected to be taken into account by employers when hiring workers. The same holds true in the case of bank managers, security officers, and other agents who make selective choices that impose nonnegligible burdens or disadvantages on people who share certain tangible features irrespective of whether or not they belong to the class with suspect grounds of discrimination. They care about certain features, ethnicity and gender for example, or, for that matter, age, education, and sartorial appearance, because they care about other characteristics that can only be ascertained indirectly.

Statistical discrimination is selection by means of tangible characteristics that function as proxies for intangibles. It operates on profiles of types of persons that support expectations about their dispositions and future behavior. A profile is a set of generic characteristics which in conjunction support a prediction that a person who fits the profile also exhibits other characteristics which are not yet manifest. A statistically sound profile is a profile that supports this prediction by faultless statistical reasoning. Technically speaking, profiles are conditional probabilities. They assign a probability estimate α to a person (i) who has a certain intangible behavioral trait (G) on the condition that they are a person of a certain type (F) with specific characteristics (F', F'', F''', ...).

$$p(G_i|F_i) = \alpha$$ [28]

The practice of profiling or making selective choices by proxy (i.e. the move from one set of personal features to another set of personal features based on a statistical correlation) is not confined to practices of illicit discrimination. It reflects a universal cognitive strategy of gaining knowledge and forming expectations not only about human beings and their behavior but about everything: observable and unobservable objects; past and future events; or theoretical entities. We move from what we believe to know about an item of consideration, or what we can easily find out about it, to what we do not yet know about it by forming expectations and making predictions. Profiling is ubiquitous also in moral reasoning and judgment. We consider somebody a fair judge if we expect fair judgments from them in the future and this expectation seems justified if they issued fair judgments in the past.

Profiling and statistical discrimination are sometimes considered dubious because they involve adverse selective choices based on personal attributes that are causally irrelevant regarding the purpose of the profiling. Ethnicity or gender, for instance, are neither causes nor effects of future economic productivity or effective leadership and, thus, may seem inappropriate criteria for hiring decisions.

Don't judge me by my color, don't judge me by my race! is a fair demand in all too many situations. Understood as a general injunction against profiling, however, it rests on a misunderstanding of rational expectations and the role of generic characteristics as predicators of personal dispositions and behavior. In the conceptual framework of probabilistic profiling, an effective predictor is a variable (a tangible personal characteristic such as age or gender) the value of which (old/young, in the middle; male/female/other) shows a high correlation with the value of another variable (the targeted intangible characteristic), the value of which it is meant to predict. Causes are reliable predictors. If the alleged cause of something were not highly correlated with it, we would not consider it to be its cause. However, good predictors do not need to have any discernible causal relation with what they are predictors for.[29]

Critical appraisals of computational profiling involve two types of misgivings. On the one hand, there are methodological flaws such as inadequate data or fallacious reasoning, on the

[28] Read "The probability that a person *i* with the characteristic F is a person who will behave in way G is α." Conditional probabilities may assign numerical probabilities ($p(G_x|F_x) = r$) or nonnumerical estimates ($p(G_x|F_x)$ is high) to intangibles. For our analysis it is irrelevant whether profiles specify numerical values, though, of course, computational profiling works with numerical values.

[29] It has long been known, for instance, that an irregular pattern of the eye-tracking movements of a person is an extremely good predictor of schizophrenia even though it is neither a cause nor a symptomatic effect of schizophrenia. See PS Holzman, LR Proctor, and DW Hughes, 'Eye-Tracking Patterns in Schizophrenia' (1973) 181 (4095) *Science* 179–181 and K Morita and other, 'Eye Movement Characteristics in Schizophrenia. A Recent Update with Clinical Implications' (2019) 40 *Neuropsychopharmacology* 2–9. The general methodological point is discussed in G Shmueli 'To Explain or to Predict?' (2010) 25 *Statistical Science* 289–310.

other hand, there are genuine moral shortcomings, for example, the lack of procedural fairness and unjust outcomes, that must be considered. Both types of misgivings are closely connected. Only sound statistical reasoning based on adequate data justifies adverse treatment which imposes nonnegligible burdens on persons, and two main causes of spurious statistics, viz. base rate fallacies and insufficiently specified reference classes, connect closely with procedural fairness.

a. Spurious Data

With regard to the informational basis of statistical discrimination, the process of specifying, collecting, and coding of relevant data may be distorted and biased in various ways. The collected samples may be too few to allow for valid generalizations or the reference classes for the data collected may be defined in inappropriate ways with too narrow a focus on a particular group of people, thereby supporting biased conjectures that misrepresent the distribution of certain personal attributes and behavioral features across different social groups. Regarding the source of the data (human behavior), problems arise because, unlike in the natural sciences, we are not dealing with irresponsive brute facts. In the natural sciences, the source of the data is unaffected by our beliefs, attitudes, and preferences. The laws of nature are independent from what we think or feel about them. In contrast, the features and regularities of human transactions and the data produced by them crucially hinge upon people's beliefs and attitudes. We act in a specific manner partly because of our beliefs about what other people are doing or intend to do, and we comply with standards of conduct partly because we believe (expressly or tacitly) that there are others who also comply with them. This affects the data basis of computational profiling in potentially unfortunate ways: prevalent social stereotypes and false beliefs about what others do or think they should do may lead to patterns of individual and social behavior which are reflected in the collected data and which, in turn, may lead to self-perpetuating and reinforcing unwanted feedback loops as described by *Noble* and others.[30]

b. Fallacious Reasoning

Against the backdrop of preexisting prejudice and bias, one may easily overestimate the frequency of unwanted behavior in a particular group and conclude that most occurrences of the unwanted behavior in the population at large are due to members of this group. There are two possible errors involved in this. Firstly, the wrong frequency estimate and, secondly, the inferential move from 'Most Fs act like Gs' to 'Most who act like Gs are Fs'. While the wrong frequency estimate reflects an insufficient data base, the problematic move rests on a base-rate fallacy, in other words, on ignoring the relative size of the involved groups.[31]

Another source of spurious statistics is insufficiently specific reference classes for individual probability estimates when relevant evidence is ignored. The degree of the correlation between two personal characteristics in a reference class may not be the same in all sub-sets of the class. Even if residence in a certain neighborhood would statistically support a bad credit rating because

[30] Biased data have found much attention in the recent literature on computational profiling. See SU Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018) and IN Cofone, 'Algorithmic Discrimination Is an Information Problem' (2019) 70 *Hasting Law Journal* 1389 for a proposal on how to deal with them and the literature referred to in the article.

[31] It does not follow from 'Most southerners are sluggards' (as northerners may feel tempted to believe) that most lazy people are southerners. There still may be more northern people that are lazy than southerners. And even if most sluggards were southerners, it would not follow that most southerners were lazy; the number of industrious southerners may still be greater.

of frequent defaults on bank loans in the area, this may not be true for a particular subgroup, for example, self-employed women living in the neighborhood for whom the frequency of loan defaults may be much lower. To arrive at valid probability estimates, we must consider all the available statistically relevant evidence and, in our example, ascertain the frequency of loan defaults for the specific reference group of female borrowers rather than for the group of all borrowers from the neighborhood. Sound statistical reasoning requires that in making probability estimates we consider all the available information and choose the maximal specific reference group of people when making conjectures about the future conduct of individuals.[32]

2. *Procedural Fairness*

Statistically sound profiles based on appropriate data still raise questions of fairness, because profiles are probabilistic and, hence, to some extent under- and over-inclusive. There are individuals with the intangible feature that the profile is meant to predict who remain undetected because they do not fit the criteria of the profile – the so-called false negatives. And there are others who do fit the profile, but do not possess the targeted feature – the so-called false positives. Under-inclusive profiles are inefficient if alternatives with a higher detection rate are available and more individuals with the targeted feature than necessary remain undetected. Moreover, false negatives undermine the procedural fairness of probabilistic profiling. Individuals with the crucial characteristic who have been correctly spotted may raise a complaint of arbitrariness if a profile identifies only a small fraction of the people with the respective feature. They are treated differently than other persons who have the targeted feature but remain undetected because they do not fit the profile. Those who have been spotted are, therefore, denied equal treatment with relevant equals. Even though the profile may have been applied consistently to all *ex-ante* equal cases – the cases that share the tangible characteristics which are the criteria of the profile – it results in differential treatment for *ex-post* equal cases – the cases that share the targeted characteristic. Because of this, selective choices based on necessarily under-inclusive profiles must appear morally objectionable.

In the absence of perfect knowledge, we can only act on what we know *ex-ante* and what we believe *ex-ante* to be fair and appropriate. Given the constraints of real life, it would be unreasonable to demand a perfect fit of *ex-ante* and *ex-post* equality. Nevertheless, a morally disturbing tension between the *ex-ante* and *ex-post* perspective on equal treatment continues to exist, and it is difficult to see how this tension could be resolved in a principled manner. Statistical profiling must be seen as a case of imperfect procedural justice which allows for degrees of imperfection and the expected detection rate of a profile should make a crucial difference for its moral assessment. A profile which identifies most people with the relevant characteristic would seem less objectionable than a profile which identifies only a small number. All profiles can be procedurally employed in an *ex-ante* fair way, but only profiles with a reasonably high detection deliver *ex-post* substantive fairness on a regular basis and can be considered procedurally fair.[33]

[32] This is the requirement of *Carnap's Principle of Total Evidence* (R Carnap, *Logical Foundations of Probability* (1950) 211). For the principle of maximally specific reference classes see C Hempel, *Aspects of Scientific Explanation and Other Essays* (1965), ch. 3.4. Meeting *Carnap's* principle and, therefore, choosing the most specific reference class that makes a statistic difference to arrive at valid probability estimates for individuals is, as we shall see in the next section, not just a requirement of epistemic rationality but also of procedural fairness. It is necessary to steer clear of avoidable over-inclusiveness (false positives) and to protect individual persons from substantively unjust treatment.

[33] Note, however, that there is not a uniquely adequate and incontestable way to fix the idea of a reasonably high detection rate. What is judged as reasonable also hinges upon the respective assessments of available alternative procedures.

Let us turn here to over-inclusiveness as a cause of moral misgivings. It may be considered unfair to impose a disadvantage on somebody for the only reason that they belong to a group of people most members of which share an unwanted feature. Over-inclusiveness means that not all members of the group share the targeted feature as there are false positives. Therefore, fairness to individuals seems to require that every individual case should be judged on its merits and every person on the basis of features that they actually have and not on merely predictable features that, on closer inspection, they do not have. Can it ever be fair, then, to make adverse selective choices based on profiles that are inevitably to some extent over-inclusive?

To be sure, *Don't judge me by my group!* is a necessary reminder in all too many situations, but as a general injunction against profiling it is mistaken. It rests on a distorted classification of allegedly different types of knowledge. Contrary to common notions, there is no categorical gap between statistical knowledge about groups of persons and individual probability estimates derived from it, on the one hand, and knowledge about individuals that is neither statistical in nature nor probabilistic, on the other. What we believe to know about a person is neither grounded solely on what we know about that person as a unique individual at a particular time and place nor independent of what we know about other persons. It is always based on information that is statistical in nature about groups of others who share or do not share certain generic features with them and who regularly do or do not act in similar ways. Our knowledge about persons and, indeed, any empirical object consists in combinations of generic features that show some stability over time and across a variety of situations. *Don't judge me by my group* thus, leads to *Don't judge me by my past*. Though not necessarily unreasonable, both demands cannot be strictly binding principles of fairness: *Do not judge me and do not develop expectations about me in the light of what I was or what I did in the past and what similar people are like in the past and present* cannot be reasonable requests.

As a matter of moral reasoning, we approve of or criticize personal dispositions and actions because they are dispositions or actions of a certain type (e.g. trustworthiness or lack thereof, promise keeping or promise breaking) and not because they are dispositions and actions of a particular individual. The impersonal character of moral reasons and evaluative standards is the very trademark of morality. Moral judgments are judgments based on criteria that equally apply to all individuals and this presupposes that they are based on generic characterizations of persons and actions. If the saying *individuum est ineffabile* were literally true and no person could be adequately comprehended in terms of combinations of generic characterizations, the idea of fairness to individuals would become vacuous. Common standards for different persons would be impossible.

We may still wonder whether adverse treatment based on a statistically sound profile is fair if it were known or could easily be known that the profile, in the case of a particular individual, does not yield a correct prediction. *Aristotle* discussed the general problem involved here in book five of his Nicomachean Ethics. He conceived of justice as a disposition to act in accordance with law-like rules of conduct that in general prescribe correct conduct but nevertheless may go wrong in special cases. *Aristotle* introduces the virtue of equity to compensate for this shortcoming of rule-governed justice. Equity is the capacity which enables an agent to make appropriate exemptions from established rules and to act on what are the manifest merits of an individual case. The virtue of equity, *Aristotle* emphasized, does not renounce justice but achieves 'a higher degree of justice'.[34] *Aristotle* conceives of equity as a remedial virtue that improves on the unavoidable imperfections of rule-guided decision-making. This provides a suitable starting

[34] Aristotle, *Nicomachean Ethics* (Fourth century BC) NE 1137b.

point for a persuasive answer to the problem of manifest over-inclusiveness. In the absence of fuller information about a person, adverse treatment based on a statistically sound profile may reasonably be seen as fair treatment, but it may still prove unfair in the light of fuller information. Fairness to individuals requires that we do not act on a statistically sound profile in adversely discriminatory ways if we know (or could easily find out) that the criteria of the profile apply but do not yield the correct result for a particular individual.[35]

3. *Measuring Fairness*

Statistical discrimination by means of computational profiling is not necessarily morally objectionable or unfair if it serves a legitimate purpose and has a sound statistical basis. The two features of probabilistic profiles that motivate misgivings, over-inclusiveness and under-inclusiveness, are unavoidable traits of human cognition and evaluation in general. They, therefore, do not justify blanket condemnation. At the same time, both give reason for moral concern.

Statisticians measure the accuracy of predictive algorithms and profiles in terms of sensitivity and specificity. The sensitivity of a profile measures how good it is in identifying true positives, individuals who fit the profile and who do have the targeted feature; specificity measures how effective it is in avoiding false positives, individuals who fit the profile but do not have the targeted feature. If the ratio of true positives to false negatives of a profile (sensitivity) is low, under-inclusiveness leads to procedural injustice. Persons who have been correctly identified by the profile may complain that they have been subjected to an arbitrarily discriminating procedure because they are not receiving the same treatment as those individuals who also have the targeted feature but who, due to the low detection rate, are not identified. This is a complaint of procedural but not of substantive individual injustice as we assume that the person has been correctly identified and, indeed, has the targeted feature. In contrast, if the ratio of false positives to the true negatives of a profile (specificity), is high, over-inclusiveness leads to procedural as well as to substantive individual injustice because a person is treated adversely for a reason that does not apply to that individual. A procedurally fair profile is, therefore, a profile that minimizes the potential unfairness which derives from its inevitable under- and over-inclusiveness.

Note the different ways in which base-rate fallacies and disregard for countervailing evidence relate to concerns of procedural fairness. Ignoring evidence leads to over-estimated frequencies of unwanted traits in a group and to unwarranted high individual probability-estimates, thereby increasing the number of false positives, in other words, members of the respective group who are wrongly expected to share it with other group members. In contrast, base-rate fallacies do not raise the number of false positives but the number of false negatives. By themselves, they do not necessarily lead to new cases of substantive individual injustice, (i.e. people being treated badly because of features which they do not have). The fallacy makes profiling procedures less

[35] This is just another application of *Carnap's* principle of total evidence and the requirement of maximally specific reference classes, in this case a class with only one known element. There are casuistic considerations that make the Aristotelian plea of equitable judgment and the demand of individual fairness less stringent than it may appear. There is no unambiguous way to decide what can be 'easily known' about a person; and there are limits to what may be morally required (or permissible) to obtain fuller personal information. There also may be unwanted external effects. If it is known that officials do allow for 'special cases', doubts as regarding the impartial application of profiles may come up; moreover, people may come to believe (perhaps wrongly) that they also will be given an exemption and not be treated in accordance with the profile, underestimating existing risks. It is difficult, however, to substantiate considerations of this kind and their relative weight will easily be overrated compared with the weight of individual fairness. Cf. for a different assessment of considering individual cases on their merits: F Schauer, *Profiles, Probabilities, and Stereotypes* (2003) ch. 8.

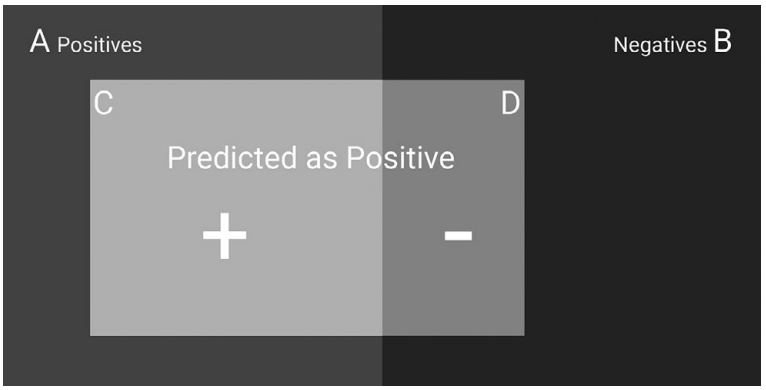

FIGURE 14.1 Fairness-index for statistical profiling based on measures for the under- and over-inclusiveness of profiles (The asymmetry of the areas C and D is meant to indicate that we reasonably expect statistical profiles to yield more true than false positives.)

efficient than they could be if base-rates were properly accounted for and, at the same time, also leads to objectionable discrimination because the false negatives are treated better than the correctly identified negatives.[36]

Our discussion suggests the construction of a fairness-index for statistical profiling based on measures for the under- and over-inclusiveness of profiles. For the sake of convenience, let us assume (a) a fixed set of individual cases that are subjected to the profiling procedure ($A \cap B$) and (b) a fixed set of (true or false) positives ($C \cap D$, see Figure 14.1). Let us further define 'sensitivity' as the ratio of true positives to false negatives and 'specificity' as the ratio of false positives to true negatives.

The sensitivity of a profile will then equal the ratio |C| / |A| and since |C| may range from 0 to |A|, sensitivity will range between 0 and 1 with 1 as the preferred outcome. The specificity of a profile will equal the ratio |D| / |B| and since |D| may range from 0 to |B| specificity will range between 0 and 1, this time with 0 as the preferred outcome. The overall statistical accuracy of a profile or algorithm could then initially be defined as the difference between the two ratios which range between –1 and +1.

$$-1 < |C|/|A| \text{ minus } |D|/|B| < +1$$

This would express, roughly, the intuitive idea that improving the statistical accuracy of a profile means maximizing the proportion of true, and minimizing the proportion of false, positives.[37]

It may seem suggestive to define the procedural fairness of a profile in terms of its overall statistical accuracy because both values are positively correlated. Less overall accuracy means more false negatives or positives and, therefore, less procedural fairness and more individual injustice. To equate the fairness of profiling with the overall statistical accuracy of the profile

[36] To assess 'algorithmic injustice' fairly, moral assessments of discriminating practices must be based on judgments of comparative and not of noncomparative (absolute) justice. If the purpose of a practice is legitimate and the burdens involved are reasonable, the crucial question is not whether it leads to wrong decisions in individual cases but how it compares in this regard with alternative practices that serve the same purpose and involve similar burdens.

[37] This is meant as a sketch to illustrate what is involved in the idea of a fairness-index for profiles based on ideas of statistical accuracy. An advanced index may involve a more sophisticated conception of overall statistical accuracy, which like the *Receiver Operator Characteristic* (ROC) familiar from the methodology of statistical measurement, does not work on binary measurements of true or false positives but on numerical probabilities estimates for individuals. Clearly these questions require more inquiry and reflection.

implies that false positive and false negatives are given the same weight in the moral assessment of probabilistic profiling, and this seems difficult to maintain. If serious burdens are involved, we may think that it is more important to avoid false positives than to make sure that no positives remain undetected. It may seem more prudent to allow guilty parties to go unpunished than to punish the innocent. In other cases, with lesser burdens for the adversely affected and more serious benefits for others, we may think otherwise: better to protect some children who do not need protection from being abused, than not to protect children who urgently need protection.

Two conclusions follow from these observations about the variability of our judgments concerning the relative weight of true and false positives for the moral assessment of profiling procedures by a fairness-index. Firstly, we need a weighing factor β to complement our formula for overall statistical accuracy to reflect the relative weight that sensitivity and specificity are supposed to have for adequate appraisal or the procedural fairness of a specific profile.

$$\beta \times |C|/|A| \text{ minus } |D|/|B|$$

Secondly, because the value of β is meant to reflect the relative weight of individual benefits and burdens deriving from a profiling procedure, not all profiles can be assessed by means of the same formula because different values for β will be appropriate for different procedures. The nature and significance of the respective benefits and burdens is partly determined by the purpose and operationalization of the procedure and partly a matter of contingent empirical conditions and circumstances. The value of β must, therefore, be determined on a case-by-case basis as a matter of securing comparative distributive justice among all persons who are subjected to the procedure in a given setting.

IV. CONCLUSION

The present discussion has shown that the moral assessment of discriminatory practices is a more complicated issue than the received understanding of discrimination allows for. Due to its almost exclusive focus on supposedly illicit grounds of unequal treatment, the received understanding fails to provide a defensible account of how to distinguish between selective choices which track generic features of persons that are morally objectionable and others that are not.

It yields verdicts of wrongful discrimination too liberally and too sparingly at the same time: too liberally, because profiling algorithms such as the Allegheny Family Screening Tool (AFST) discussed by *Virginia Eubanks* in her *Automating Inequality* that work on great numbers of generic characteristics can hardly be criticized as being unfairly discriminating for the only reason that ethnicity and income figures among the variables make a difference for the identification of children at risk. It yields verdicts too sparingly because a limited list of salient characteristics and illicit grounds of discrimination is not helpful in the identifying of discriminated groups of persons who do not fall into one of the familiar classifications or share a salient set of personal features.

For the moral assessment of computational profiling procedures such as the Allegheny Algorithm, it is only of secondary importance whether it employs variables that represent suspect characteristics of persons, such as ethnicity or income, and whether it primarily imposes burdens on people who share these characteristics. If the algorithm yields valid predictions based on appropriately collected data and sound statistical reasoning and if it has a sufficiently high degree of statistical accuracy, the crucial question is whether the burdens it imposes on some people are not unreasonable and disproportional and can be justified by the benefits that it brings either to all or at least to some people.

The discriminatory power and the validity of profiles is for the most part determined by their data basis and by the capacity of profiling agents to handle heterogeneous information about persons and generic personal characteristics to decipher stable patterns of individual conduct from the available data. The more we know about a group of people who share certain attributes, the more we can learn about the future behavior of its members. Further, the more we know about individual persons, the more we are able to know more about the groups to which they belong.[38] Profiles based on single binary classifications, for instance, male or female, native or alien, Christian or Muslim, are logically basic (and ancient) and taken individually offer poor guidance for expectations. Valid predictions involve complex permutations of binary classifications and diverse sets of personal attributes and features. Computational profiling with its capacity to handle great numbers of variables and possibly with online access to a vast reservoir of data is better suited for the prediction of individual conduct than conventional human profiling based on rather limited information and preconceived stereotypes.[39]

Overall, computational profiling may prove less problematic than conventional stereotyping or old-fashioned statistical profiling. Advanced algorithmic profiling enhanced by AI is not a top-down application of a fixed set of personal attributes to a given set of data to yield predictions about individual behavior. It is a self-regulated and self-correcting process which involves an indefinite number of variables and works both from the top down and the bottom up, from data mining and pattern recognition to the (preliminary) definition of profiles and from preliminary profiles back to data mining, cross-checking expected outcomes against observed outcomes. There is no guarantee that these processes are immune to human stereotypes and void of biases, but many problems of conventional stereotyping can be avoided. Ultimately, computational profiling can process indefinitely more variables to predict individual conduct than conventional stereotyping and, at the same time, draw on much larger data sets to confirm or falsify predictions derived from preliminary profiles. AI and data mining via the Internet, thus, open the prospect of a more finely grained and reliable form of profiling, thereby overcoming the shortcomings of conventional intuitive profiling. On that note, I recall a colleague in Shanghai emphasizing that he would rather be screened by a computer program to obtain a bank loan than by a potentially ill-informed and corrupt bank manager.

[38] As a rule of thumb, this seems to be true, even if it is kept in mind that more information normally also means more irrelevant information. There is not only the problem of knowing too little about persons to make valid predictions. There is also the problem of knowing too much about the individual case and the need to suppress the "noise" of irrelevant information to discern stable patterns of behavior. Sorting out relevant information, however, typically requires even more information. For an accessible account of noise and over-fittingness see D Spiegelhalter, *The Art of Statistics. Learning from Data* (2019) chapter 6.

[39] For a more skeptical assessment of Big Data and the advances of scientific prediction by means of machine learning cf. S Succi and PV Coveney, 'Big Data: the End of the Scientific Method?' (2019) A 377 *Philosophical Transactions Royal Society* 20180145.

15

Discriminatory AI and the Law

Legal Standards for Algorithmic Profiling

Antje von Ungern-Sternberg

I. INTRODUCTION

One of the great potentials of Artificial Intelligence (AI) lies in profiling. After sifting through and analysing huge datasets, intelligent algorithms predict the qualities of job candidates, the creditworthiness of potential contractual partners, the preferences of internet users, or the risk of recidivism among convicted criminals. However, recent studies show that building and applying algorithms based on profiling can have discriminatory effects. Hiring algorithms may be biased against women,[1] and credit rating algorithms may disfavour people living in poorer neighbourhoods.[2] Algorithms can set prices or convey information to internet users classified by gender, race, sexual orientation, or disability,[3] and predicting recidivism algorithmically can have a disparate impact on people of colour.[4]

While some observers stress the particular danger posed by discriminatory AI,[5] others hope that it might eventually end discrimination[6]. Before examining the particular challenges of discriminatory AI, one should keep in mind that human decision-making is also affected by prejudices and stereotypes, and that algorithms might help avoid and detect manifest and hidden forms of discrimination. Nevertheless, possible discriminatory effects of AI need to be assessed for several reasons. First, algorithms can perpetuate existing societal inequalities and stereotypes if they are trained with datasets that reflect inequalities and stereotypes. Second, algorithms used

[1] C O'Neil, *Weapons of Math Destruction* (2017) (hereafter 'O'Neil, *Weapons of Math Destruction*') 105 *et seq*; P Kim, 'Data-Driven Discrimination at Work' (2017) 58 *William & Mary Law Review* 857, 869 *et seq*.

[2] O'Neil, *Weapons of Math Destruction* (n 1) 141 *et seq*; J Allen 'The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining' (2019) 46 *Fordham Urban Law Journal* 219.

[3] J Angwin and T Parris, 'Facebook Lets Advertisers Exclude Users by Race' (*ProPublica*, 28 October 2016) www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race; A Kofman and A Tobin, 'Facebook Ads Can Still Discriminate against Women and Older Workers, Despite a Civil Rights Settlement' (*ProPublica*, 13 December 2019) www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement; N Kayser-Bril, 'Automated Discrimination: Facebook Uses Gross Stereotypes to Optimize Ad Delivery' (*AlgorithmWatch*, 18 October 2020) https://algorithmwatch.org/en/story/automated-discrimination-facebook-google/; S Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (2020) 35 *Berkeley Technology Law Journal* 367 (hereafter Wachter, 'Affinity Profiling').

[4] J Angwin, J Larson, S Mattu, and L Kirchner, 'Machine Bias' (*ProPublica* 23 May 2016) www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (hereafter Angwin and others, 'Machine Bias').

[5] O'Neil, *Weapons of Math Destruction* (n 1). *Cf*. also K Zweig, *Ein Algorithmus hat kein Taktgefühl* (3rd ed., 2019) 211.

[6] J Kleinberg and others, 'Discrimination in the Age of Algorithms' (2018) 10 *Journal of Legal Analysis* 1 (hereafter Kleinberg and others, *'Discrimination in the Age of Algorithms*').

by large companies or state agencies affect many people. Third, the discriminatory effects of AI have not been easy to detect and to prove until now. What's more, some of the predictions resulting from AI analysis cannot be verified. If a person does not obtain credit, then she can hardly prove creditworthiness; likewise, if an applicant is not hired, there is no way he can prove to be a good employee. Finally, algorithms are often perceived as particularly rational or neutral, which may prevent questioning of its results.

Therefore, this article offers an assessment of the legality of discriminatory AI. It concentrates on the question of material legality, leaving many other important issues aside, namely the crucial question of detecting and proving discrimination.[7] Drawing on legal scholarship showing discriminatory effects of AI,[8] this article analyses existing norms of anti-discrimination law,[9] depicts the role of data protection law,[10] and treats suggested standards such as a right to reasonable inferences[11] or 'bias transforming' fairness metrics that help secure substantive rather than mere formal equality.[12] This chapter shows that existing standards of anti-discrimination law already imply how to assess the legality of discriminatory effects, even though it will be helpful to develop and establish these aspects in more detail. As this assessment involves technical and legal questions, both lawyers as well as data and computer scientists need to cooperate. This article proceeds in three steps. After explaining the legal framework for profiling and automated decision-making (II), the article analyses the different causes for discrimination (III) and develops the relevant aspects of a legality or illegality assessment (IV).

[7] Some of the arguments developed in this chapter can also be found in A von Ungern-Sternberg, 'Diskriminierungsschutz bei algorithmenbasierten Entscheidungen' in A Mangold and M Payandeh (ed), *Handbuch Antidiskriminierungsrecht – Strukturen, Rechtsfiguren und Konzepte* (forthcoming) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3828696.

[8] *Cf.* n 1–6; B Friedman and H Nissenbaum, 'Bias in Computer Systems' (1996) 14 *ACM Transactions on Information Systems* 330(333 *et seq*) (hereafter Friedman and Nissenbaum, 'Bias in Computer Systems'); Calders and I Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures' in B Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 43, 50 *et seq* (hereafter Calders and Žliobaitė, 'Unbiased Computational Processes'); S Barocas and A Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671, 681 *et seq* (hereafter Barocas and Selbst, 'Big Data's Disparate Impact'); C Orwat, *Diskriminierungsrisiken durch Verwendung von Algorithmen* (*Antidiskriminierungsstelle des Bundes*, 2019) 34 *et seq*, 77 *et seq* www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.html (hereafter Orwat, *Diskriminierungsrisiken*).

[9] P Hacker, 'Teaching Fairness to Artificial Intelligence' (2018) 55 *Common Market Law Review* 1143; F Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 *The International Journal of Human Rights* 1572; J Gerards and F Zuderveen Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (*SSRN*, 2020) https://ssrn.com/abstract=3723873 (hereafter Gerards and Zuderveen Borgesius, 'Protected Grounds'); Wachter, 'Affinity Profiling' (n 3); S Wachter, B Mittelstadt and C Russell, 'Why Fairness Cannot Be Automated' (*SSRN*, 2020) https://ssrn.com/abstract=3547922 (hereafter Wachter, Mittelstadt and Russell, 'Why Fairness Cannot Be Automated'); M Martini, *Blackbox Algorithmus: Grundfragen einer Regulierung Künstlicher Intelligenz* (2019) 73-91, 230-249.

[10] W Schreurs, M Hildebrandt, E Kindt, and M Vanfleteren, 'Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector' in M Hildebrandt and S Gutwirth, *Profiling the European Citizen* (2008) 241 (hereafter Schreurs and others, *Profiling*); I Cofone, 'Algorithmic Discrimination Is an Information Problem' (2019) 70 *Hastings Law Journal* 1389, 1416 *et seq* (hereafter Cofone, 'Algorithmic Discrimination'); S Wachter and B Mittelstadt, 'A Right to Reasonable Inferences' (2019) *Columbia Business Law Review*, 494 (hereafter Wachter and Mittelstadt, 'A Right to Reasonable Inferences'); A Tischbirek, 'Artificial Intelligence and Discrimination' in T Wischmeyer and T Rademacher (eds), *Regulation Artificial Intelligence* (2020) 104.

[11] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

[12] S Wachter, B Mittelstadt and C Russell, 'Bias Preservation in Machine Learning' *West Virginia Law Review* (forthcoming) https://ssrn.com/abstract=3792772 (hereafter Wachter, Mittelstadt and Russell, 'Bias Preservation').

II. LEGAL FRAMEWORK FOR PROFILING AND DECISION-MAKING

Using AI to profile involves different steps for which different legal norms apply. A legal definition of profiling can be found in the General Data Protection Regulation (GDPR). It 'means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.[13] Thus, profiling describes an automated process (as opposed to human instances of profiling, for instance by a police profiler) affecting humans (as opposed to AI optimising machines, for example) which increasingly relies on AI for detecting patterns, establishing correlations, and predicting human characteristics. Without going into detail about different possible definitions of AI,[14] profiling algorithms qualify as 'intelligent' as they can solve a defined problem, in other words, they can make predictions about unknown facts based on an analysis of data and patterns. After obtaining the profiling results on characteristics such as credit risk, job performance, or criminal behaviour, machines or humans may then make decisions on loans, recruiting, or surveillance. Thus, it is helpful to distinguish between (1) profiling and (2) decision-making. One can broadly assume that anti-discrimination law governs decision-making, whereas data protection law governs the input of personal data needed for profiling. A closer look reveals, however, that things are more complex than that.

1. *Profiling*

The process of profiling is comprised of several steps. The first step involves collecting data for training purposes. The second step entails building a model for predicting a certain outcome based on particular predictors (using a training algorithm). The final step applies this model to a particular person (using a screening algorithm).[15] Generally speaking, the first and the third steps are governed by data protection law because they involve the processing of personal data – either for establishing the dataset or for screening and profiling a particular person. The GDPR covers the processing of personal data by state actors and state parties alike, and requires that processing is based on the consent of the data subject or on another legal ground. Legal grounds can include necessary processing for the performance of a contract, compliance with a legal obligation, or for the purposes of legitimate interests.[16] Furthermore, the Law Enforcement Directive (LED) provides that the processing of personal data by law enforcement authorities must be necessary for preventing and prosecuting criminal offences or executing criminal penalties.[17] Thus, data protection law requires a sufficient legal basis for collecting and processing training data, as well as for collecting and processing the data of a specific person being profiled. Public authorities will mostly rely on statutes, while private companies will often rely on the necessity for the performance of a contract or base their activities on legitimate interests

[13] GDPR, Article 4(4).
[14] S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (4th ed. 2022) 19–23.
[15] Schreurs and others, *Profiling* (n 10) 246; Kleinberg and others, *'Discrimination in the Age of Algorithms'* (n 6) 22.
[16] GDPR, Article 6(1)(a)(b)(c)(e) or (f). According to Article 2(2), the GDPR only applies to the processing of data 'by automated means' or if it forms part of a 'filing system' or is intended to form part of such a system. Thus, algorithmic (i.e. automated) forms of profiling fall under this heading.
[17] Article 8(1) Directive (EU) 2016/680 (LED). The GDPR does not apply to these activities of law enforcement authorities, *cf.* GDPR, Article 2 (2)(d).

or the consent of the data subjects. The processing of special ('sensitive') data, including personal data revealing racial or ethnic origin, political opinion, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, and data concerning health or data concerning a natural person's sex life or sexual orientation, must comply with additional legality requirements.[18]

Yet, several questions remain. First, the second step, building the profiling model, is not covered by data protection law if the data is anonymised. Data protection law only applies to personal data, i.e. information relating to an identified or identifiable natural person.[19] Since it is not necessary to train a profiling algorithm on personalised data, datasets are regularly anonymised before the second step.[20] Some authors suggest that data subjects whose personal data have been collected during the first step should have the right to object to anonymisation, as this also constitutes a form of data processing.[21] However, even if this right exists for those cases when processing is based on consent, data subjects might not bother to object. Subjects may not bother to object either because they benefit from data collection, as in participating in a supermarket's consumer loyalty programme or internet web page access in exchange for accepting cookies, or because they are not immediately affected by the profiling. It is important to keep in mind that the data subjects providing training data (step one) may be completely different from the data subjects which are later profiled (step three).

Second, even during the first and the third step, it is not always clear whether personal data is being processed. Big data analysis can refer to all kinds of data. In a supermarket, for example, shopping behaviour can correlate not only with the date and time of shopping, but also with the contents and the movements (speed, route) of the shopping trolley. In an online environment, data ranging from online behaviour to keystroke patterns and the use of a certain end device may be linked to characteristics like price-sensitivity or creditworthiness. In this context, singling out a person as an individual, even if the data controller does not know the individual's name, should be enough to consider a person 'identifiable'.[22] Thus, cases where a company can recognise and trace an individual consumer or where a state agency can single out an individual fall under data protection law.

Third, it is disputed how the methodology of profiling and the profiling result (i.e. the profile of a particular person) should be treated in data protection law. It is helpful to distinguish different categories of data, notably collected data, like data submitted by the data subject or observed by the data controller, and data inferred from collected data, such as profiles.[23]

[18] GDPR, Article 9; LED, Article 10.

[19] GDPR, Article 4(1); LED, Article 3(1).

[20] Schreurs and others, *Profiling* (n 10) 248.

[21] Schreurs and others, *Profiling* (n 10) 248–253.

[22] *Cf.* that GDPR, Article 4(1) and LED, Article 3(1) also refer to an 'online identifier'; D Korff, 'New Challenges to Data Protection Study – Working Paper No 2: Data Protection Laws in the EU: The Difficulties in Meeting the Challenges Posed by Global Social and Technical Developments' (*European Commission DG Justice, Freedom and Security Report* 15 January 2010) https://ssrn.com/abstract=1638949, 45–48 (hereafter Korff, 'New Challenges to Data Protection'); Schreurs and others, *Profiling* (n 10) 247; F Zuiderveen Borgesius, 'Singling Out People without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation' (2016) 32 *Computer Law & Security Review* 256; F Zuiderveen Borgesius and J Poort, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40 *Journal of Consumer Policy* 347 (356–358).

[23] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling' WP251rev.01 (*Directorate C of the European Commission*, 6 February 2018) 8 https://ec.europa.eu/newsroom/article29/document.cfm?doc_id=49826; Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10) 516; R Broemel and H Trute, 'Alles nur Datenschutz' (2016) 27 *Berliner Debatte Initial* 50 (52).

Even though it is misleading to qualify inferred data as 'economy class' data,[24] inferred data is different from collected data in two regards. First, the methodology of inference varies considerably. Based on collected data, physicians diagnose medical conditions, lawyers assess the legality of acts, professors evaluate exams, journalists judge politicians, economists predict the behaviour of consumers, and internet users rate the service of online-sellers, each according to different scientific or value-based standards. Furthermore, one has to acknowledge that the inference itself is an accomplishment based on effort, values, qualifications, and/or skills. Profiling (i.e. algorithmic inferences about humans), also exhibits these two characteristics. Its distinct methodology is determined by its training and profiling algorithms, and its achievement is legally recognised, for example, by intellectual property protecting profiling algorithms[25] or by other rights like freedom of speech.[26]

This does not imply that predictions about characteristics and qualities of a particular person do not qualify as personal data. The Article 29 Data Protection Working Party, the precursor of today's European Data Protection Board, specified that data related to an individual if the data's content, result, or purpose was sufficiently linked to a particular person.[27] If a person's profile provides information about her (content), if it aims to evaluate her (purpose), and if using the profile will likely have an impact on her rights and interests (result), then the profile must be considered personal data.[28] However, the characteristics of inferred data can have an impact upon the data subject's rights. Notably, the right to rectification of inaccurate personal data[29] only refers to instances of inaccuracy which can be verified (e.g. the attribution of collected or inferred data to the wrong person). But the right generally does not include the appropriate (medical, legal, economic, et cetera) methodology of inferring information, as this is beyond the reach of data protection law.[30] This is the reason why scholars call for a right to reasonable interferences.[31] Yet, one might argue that profiling, as opposed to other methods of inferring data, is indeed, at least partially, regulated by data protection law.[32] In any event, profiling is not an activity privileged by the GDPR. The GDPR clauses promoting data processing for 'statistical purposes'[33] are not intended to facilitate profiling.[34] This follows from the wording of the

[24] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10) 494.

[25] GDPR, Recital 63; *cf.* BGHZ 200, 38 (BGH VI ZR 156/13) on the trade secret of Schufa, the German (private) General Credit Protection Agency, concerning its scoring algorithm.

[26] J Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 *UC Davis Law Review* 1149; note that GDPR, Article 85(1) demands that Member States reconcile data protection with the right to freedom of expression.

[27] Article 29 Data Protection Working Party, 'Opinion 4/2007 on the concept of personal data, 01248/07/EN WP 136' (*European Commission*, 20 June 2007) 9–12 https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf.

[28] Korff, 'New Challenges to Data Protection' (n 22) 52–53; Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10), 515–521.

[29] GDPR, Article 16; LED, Article 16.

[30] *Cf.* CJEU, Case C-434/16 *Nowak* [2017] n 52–57, on the right to rectification concerning written exams which does not extend to incorrect answers but possibly if examination scripts were mixed up by mistake.

[31] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

[32] See Section IV 3(a).

[33] GDPR, Articles 5(1)(b) and (e), 9(2)(j), 14(5)(b), 17(3)(d), 21(6), 89(1) and (2).

[34] This, however, is suggested by V Mayer-Schönberger and Y Padova, 'Regime Change? Enabling Big Data through Europe's New Data Protection Regulation' (2016) 17 *Columbia Sciences & Technology Law Review* 315 (330).

clauses, from Recital 162[35] and from the purpose of the GDPR, which is regulating profiling in order to control the risks emanating from it.[36]

2. Decision-Making

Anti-discrimination law and data protection law can govern the decisions that follow profiling.

a. Anti-Discrimination Law

Anti-discrimination provisions, grounded in national law, European Union law, and public international law, prohibit direct and (often) indirect forms of discrimination.[37] Some non-discrimination provisions address the state, while others are binding upon state and private actors. Some provisions have a closed list of protected characteristics, while others are more public.[38] Some provisions apply very broadly, covering employment or the supply of goods and services available to the public,[39] while still others have a narrower scope, merely affecting insurance contracts or management of journalistic online content, for example.[40] This chapter does not seek to examine the commonalities or differences of these provisions but rather aims to analyse if and when decision-making based on profiling may be justified.

This analysis is based on some general observations. First, anti-discrimination law applies to human and machine decisions alike. It does not presuppose a human actor. Thus, it is not relevant for anti-discrimination law whether a decision has been made solely by an algorithm, solely by a human being (based on the profile), or by both (i.e. by a human being accepting or not objecting to the decisions suggested by an algorithm). Second, anti-discrimination law distinguishes between direct and indirect discrimination, or between differential treatment and detrimental impact.[41] In EU anti-discrimination law, direct discrimination occurs when one person is treated less favourably than another is treated or would be treated in a comparable situation because of a protected characteristic such as race, gender, age, or religion.[42] Indirect

[35] '[. . .] Statistical purposes mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.'

[36] Thus, the statistical privilege is only granted if public agencies conduct statistical surveys and produce statistical results, or if similar activities take place in the public interest (and not in support of profiling a particular natural person), *cf.* J Caspar, 'Article 89' in S Simitis, G Hornung, and I Spiecker gen Döhmann (eds), *Datenschutzrecht* (2019) n 23.

[37] Article 3 German Basic Law, German General Equal Treatment Act (2006); Article 21 EU Charter of Fundamental Rights (CFA), Framework Directive 2000/78/EC, Race Directive 2000/43/EC, Goods and Services Sex Discrimination Directive 2004/113/EC, Equal Treatment Directive 2006/54/EC; Article 14 European Convention on Human Rights.

[38] For an overview see European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European Non-discrimination Law* (2010) https://fra.europa.eu/sites/default/files/fra_uploads/1510-FRA-CASE-LAW-HANDBOOK_EN.pdf; M Connolly, *Discrimination Law* (2nd ed., 2011) 15, 55, 79, 151 (hereafter Connolly, *Discrimination Law*); Gerards and Zuderveen Borgesius, 'Protected Grounds' (n 9).

[39] Article 3(1) Framework Directive 2000/78/EC; Article 3(1)(c) and (h) Race Directive 2000/43/EC; Article 3(1) Goods and Services Sex Discrimination Directive 2004/113/EC; Article 14(1) Equal Treatment Directive 2006/54/EC.

[40] In German law, §19(1) n° 2 German General Equal Treatment Act (2006) contains a specific anti-discrimination norm for private insurance contracts; §94(1) of the new State Treaty on Media (2020) forbids big media platforms to discriminate between journalistic content.

[41] *Cf.* D Schiek, 'Indirect Discrimination' in D Schiek, L Weddington, and M Bell, *Non-Discrimination Law* (2007) 323 (372) (hereafter Schiek, 'Indirect Discrimination'). This is also known as disparate treatment and disparate impact in U.S. terminology.

[42] See e.g. Framework Directive 2000/78/EC, Article 2(2)(a).

discrimination occurs when an apparently neutral provision, criterion, or practice would put members of a protected group at a particular disadvantage compared with other persons, unless this is justified.[43] Note the term 'discrimination' implies illegality in German usage, whereas differential treatment or detrimental effect can be legal if it is justified. However, this article follows the English use of the term 'discrimination' which encompasses illegal and legal forms of differential treatment or detrimental effect. Algorithmic profiling and decision-making can easily avoid direct discrimination if algorithms are prohibited from collecting or considering protected characteristics. However, if algorithms are trained on datasets reflecting societal inequalities and stereotypes (indicating, for instance, that men are better qualified for certain jobs than women), profiling and decision-making might put already disadvantaged groups (like female applicants) at a particular disadvantage. Thus, one can expect indirect discrimination to gain importance in an era of algorithmic profiling and decision-making. As a consequence, corresponding questions like "How can a particular disadvantage be established?"[44] or "What are the reasons for banning indirect discrimination?"[45] will become increasingly relevant.

Third, direct and indirect forms of discrimination, or differential treatment and detrimental effect, can be justified. Generally speaking, indirect discrimination is easier to justify than direct discrimination. In EU anti-discrimination law, indirectly causing a particular disadvantage does not amount to indirect discrimination if it 'is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary'.[46] But differential treatment can also be justified, either on narrow[47] or on broader[48] grounds, provided that it passes a proportionality test. Thus, considerations of proportionality are relevant for all attempts to justify direct and indirect forms of discrimination. This chapter submits that these considerations are significantly shaped by the commonalities of intelligent profiling and automation, as will be explained below.

b. Data Protection Law

Examining the legal framework for automated decision-making would be incomplete without Article 22 GDPR and Article 11 LED. These provisions go beyond a mere regulation of data processing by limiting the possible uses of its results. They apply to a decision 'based solely on automated processing, including profiling, which produces legal effects' concerning the data subject or 'significantly affect[ing] him or her'[49] and generally prohibit such a mode of automated decision-making unless certain conditions are met. Thus, the provisions also cover discriminatory decisions if they are automated. Furthermore, there is an explicit link between

[43] See e.g. Framework Directive 2000/78/EC, Article 2(2)(b).

[44] Wachter, Mittelstadt and Russell, 'Why Fairness Cannot Be Automated' (n 9) para V *et seq*.

[45] A Morris, 'On the Normative Foundations of Indirect Discrimination Law' (1995) 15 *Oxford Journal of Legal Studies* 199 (hereafter Morris, 'On the Normative Foundations'); C Tobler, *Limits and Potential of the Concept of Indirect Discrimination* (2008) 17–35 (hereafter Tobler, *Limits*); Connolly, *Discrimination Law* (n 38) 153–156.

[46] See e.g. Framework Directive 2000/78/EC, Article 2(2)(b)(i); Race Directive 2000/43/EC, Article 2(2)(b); Goods and Services Sex Discrimination Directive 2004/113/EC, Article 2(b); Equal Treatment Directive 2006/54/EC, Article 2(1)(b).

[47] The German Federal Constitutional Court, for example, accepted unequal treatment based on gender permissible only 'if compellingly required to resolve problems, that because of their nature, can occur only in the case of men or women' BVerfGE 85, 191 (BVerfG *1 BvR 1025/82*), Konrad-Adenauer-Stiftung, *70 Years German Basic Law* (3rd ed., 2019), 288.

[48] See e.g. Framework Directive 2000/78/EC, Articles 4 and 6; Goods and Services Sex Discrimination Directive 2004/113/EC, Article 4(5); CFR, Article 52(1) with regard to CFR, Article 21; DJ Harris and others, *Harris, O'Boyle and Warbrick: Law of the European Convention on Human Rights* (4th ed., 2018) 772–776 with regard to Art 14 ECHR (hereafter Harris and others, *European Convention on Human Rights*).

[49] GDPR, Article 22(1); LED, Article 11(1).

data protection and anti-discrimination law in Article 11 (3) LED, which prohibits profiling that results in discrimination against natural persons on the basis of special ('sensitive') data. A similar clause is missing in the GDPR, but the recitals indicate that the regulation is also intended to protect against discrimination.[50]

However, the scope and relevance of Article 22 GDPR are much debated. The courts have not yet established what 'a decision based solely on automated processing' means or what amounts to 'significant' effects.[51] Likewise, automated decision-making can still be based on explicit consent, contractual requirements, or a statutory authorisation as long as suitable measures safeguard the data subject's rights and freedoms and legitimate interests,[52] in other words, legal bases can also be understood in a restrictive or permissive way. The same applies to the anti-discrimination provision of Article 11(3) LED, which could extend to all forms, automated and human alike, of decision-making based on profiling (or be confined to automated decision-making) and which is open to different standards of scrutiny if differential treatment or factual disadvantages are justified.

3. *Data Protection and Anti-Discrimination Law*

The brief overview of relevant norms of data protection and anti-discrimination law shows that both areas of law are important in prohibiting and preventing discriminations caused by decision-making based on algorithmic profiling. Data protection law can be characterised not only as an end in and of itself, but also as a means to prevent discrimination based on data processing.[53] Such an understanding of data protection law flows from the recitals referring to discrimination,[54] from the special protection for categories of 'sensitive' data such as race, religion, political opinions, health data, or sexual orientation (which conform to the categories of protected characteristics in anti-discrimination law),[55] and from particular provisions concerning profiling.[56] These provisions do not only limit profiling and automated decision-making, but they also specify corresponding rights and duties, including rights of access ('meaningful information' about the logic of profiling),[57] rights to rectification and erasure,[58] or the duties to ensure data protection by design and by default[59] and to carry out a data protection impact assessment.[60]

[50] Recital 71 in regard to Article 22 GDPR states: '[. . .] In order to ensure fair and transparent processing in respect of the data subject, [. . .] the controller should [. . .] secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. [. . .]'. The prevention of anti-discrimination is also referred to in Recitals 75 and 85.

[51] The Article 29 Data Protection Working Party favours a broad reading of Article 22 GDPR for machine-human interaction, qualifying as automated decision-making if a human 'routinely applies automatically generated profiles to individuals', in other words, if human intervention is reduced to a mere 'token gesture'. It suggests a similarly broad understanding of significant effects, possibly including the refusal of a contract or targeted advertising; Guidelines on Automated individual decision-making (n 23) 10–11.

[52] GDPR, Articles 22(2)–(4).

[53] *Cf.* R Poscher, Chapter 16, in this volume.

[54] *Cf.* n 50 for the GDPR and LED, Recitals 23, 38, 51, and 61.

[55] GDPR, Article 9; LED, Article 10.

[56] GDPR, Article 22; LED, Article 11.

[57] GDPR, Article 15(1)(h); general information rights are granted in Articles 12–15 GDPR, Articles 12–14 LED.

[58] GDPR, Articles 16 and 17; LED, Article 16.

[59] GDPR, Article 25; LED, Article 20.

[60] GDPR, Article 35; LED, Article 27.

III. CAUSES FOR DISCRIMINATION

After examining the legal framework for profiling and decision-making, it is now crucial to ask why discrimination occurs in the context of intelligent profiling. This article suggests that one can distinguish two (partially overlapping) causes of discrimination: (1) the use of statistical correlations and (2) technological and methodological factors, commonly referred to as 'bias'.

1. *Preferences and Statistical Correlations*

American economists were the first to distinguish between taste-based discrimination and statistical discrimination ('discrimination' meaning differentiation, bearing no negative connotation). According to this distinction, discrimination either relies on preferences or implies the rational use of statistical correlations to cope with a lack of information. If, for instance, young age correlates with high productivity, a prospective employer who does not know the individual productivity of two applicants may hire the younger applicant in efforts to increase the productivity of her enterprise. Due to its rational objective, statistical discrimination seems less problematic than enacting ones' irrational preferences, for example not hiring older applicants based on a dislike for older people.[61]

It is evident that direct or indirect discrimination resulting from group profiling[62] also qualifies as statistical discrimination. Group profiling describes the process of predicting characteristics of groups, as opposed to personalised profiling which aims to identify a particular person and to predict her characteristics.[63] Data mining and automation allows for increasingly sophisticated profiles and correlations to be established. Instead of relying on a simple proxy like age, gender, or race, decision-making can now be based on a complex profile. The use of these profiles rests on the assumption that the members of a certain group defined by specific data points also exhibit certain (unknown, but relevant) characteristics. Examples of this practice can be found everywhere as more and more private companies and state agencies use algorithmic group profiles. Companies, for example, rely on group profiles assessing the capabilities of prospective employees, the risks of prospective insurees, or the preferences of online consumers. But state agencies also take group profiles into account, when, for instance, predicting the inclination to commit an offence or the need for social assistance.[64]

Even if contrasted with taste-based discrimination, statistical discrimination is not wholly unproblematic. Sometimes, it implies direct discrimination based on protected characteristics, for example if certain risks allegedly correlate with race, religion, or gender.[65] Furthermore, statistical discrimination means that the predicted characteristic of a group is attributed to its

[61] E Phelps, 'The Statistical Theory of Racism and Sexism' (1972) 62 *The American Economic Review* 659; *cf.* G Britz, *Einzelfallgerechtigkeit versus Generalisierung* (2008) 15 *et seq* (hereafter Britz, *Einzelfallgerechtigkeit*). The term statistical discrimination should not be confused with the statistical proof of (indirect) discrimination.

[62] The term 'profiling' means 'group profiling' unless otherwise noted.

[63] M Hildebrandt, 'Defining Profiling: A New Type of Knowledge' in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen* (2008)17, 20–23 (hereafter Hildebrandt, 'Defining Profiling').

[64] On predictive policing based on group profiles see E Joh, 'The New Surveillance Discretion' (2016) 15 *Harvard Law & Policy Review* 24; A Ferguson, 'Policing Predictive Policing' (2017) 94 *Washington University Law Review* 1109, 1137–1143; examples of European state practice can be found in AlgorithmWatch, 'Automating Society' (*Algorithm Watch*, January 2019) https://algorithmwatch.org/en/automating-society/; e.g. in employment service 43, 108, 121, in children and youth assistance and protection 50, 61, 101, 115, in health care 88–89.

[65] A von Ungern-Sternberg, 'Religious Profiling, Statistical Discrimination and the Fight against Terrorism in Public International Law' in R Uerpmann-Wittzack, E Lagrange and S Oeter (eds), *Religion and International Law* (2018), 191 (hereafter Ungern-Sternberg, 'Religious Profiling').

members, even though there is only a certain probability that a group member shares this characteristic[66] and even though the attributes themselves might be negative (e.g. a correlation of race and delinquency or of age and mental capacity).[67]

Finally, it should be noted that discrimination can be based on a combination of taste and statistical correlations. This is the case, for example, when companies take into account consumer preferences predicted from group profiles. Online platforms respond to presumed user preferences when displaying news, search results, or information on prospective employers, dates, or goods. This can also raise problems. Predicting group preferences might disadvantage certain groups of users, like female or Black jobseekers who are shown less attractive job offers than White male men.[68] Additionally, group preferences might be discriminatory and lead to discriminatory decisions. Google searches for Black Americans might yield ads for criminal record checks, the comments of people of colour or homosexuals might be less visible on online platforms, and dating platform users might be categorised along racial or ethnic lines.[69]

2. *Technological and Methodological Factors*

Discrimination based on correlations can also entail (further) disadvantages and biases stemming from the profiling method. In the literature, this phenomenon is sometimes called 'technical bias'.[70] This term can be misleading, however, as these biases also occur in the context of human profiling.[71] Furthermore, these biases result not only from technical circumstances, but also from deliberate methodological decisions. These decisions involve collecting the training data (step 1), specifying a concrete outcome to predict (including one or several target variables indicating this outcome) (step 2), choosing possible predictor variables that are made available to the training algorithm (step 3), and finally, after the training algorithm has chosen and assessed the relevant predictor variables for the predicting model (i.e. after building the screening algorithm) validating the screening algorithm in another (verification) dataset (step 4).[72] All of these decisions can involve biases.

a. Sampling Bias

A sampling bias may follow from unrepresentative datasets that are used to train (step 1) and to validate (step 4) algorithms.[73] Transferring the result of machine learning to new data rests on the assumption that this new data has similar characteristics as the dataset used to train and

[66] This is why Hildebrandt (in Hildebrandt, 'Defining Profiling' (n 63) 21) considers group profiles 'non-distributive profiles'.

[67] On this see Britz, *Einzelfallgerechtigkeit* (n 61) 23.

[68] T Speicher and others, 'Potential for Discrimination in Online Targeted Advertising' (2018) 81 *Proceedings of Machine Learning research* 1.

[69] L Sweeney, 'Discrimination in Online Ad Delivery' (2013) 56 *Communications of the ACM* 44; N Kayser-Bril, 'Automated Moderation Tool from Google Rates People of Color and Gays as "Toxic"' (*Algorithmwatch*, 19 May 2020) https://algorithmwatch.org/en/story/automated-moderation-perspective-bias/; J Hutson and others, 'Debiasing Desire: Addressing Bias & Discrimination on Intimate Platforms' (2018) 2 *Proceedings of the ACM on Human-Computer Interaction* 1.

[70] There does not seem to be an established terminology yet, *cf*. Friedman and Nissenbaum, 'Bias in Computer Systems' (n 8) 333; Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 50; Barocas and Selbst, 'Big Data's Disparate Impact' (n 8) 681.

[71] Britz, *Einzelfallgerechtigkeit* (n 61) 18–22.

[72] Kleinberg and others, '*Discrimination in the Age of Algorithms*' (n 6) 22.

[73] Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 51; Barocas and Selbst, 'Big Data's Disparate Impact (n 8) 684; Orwat, *Diskriminierungsrisiken* (n 8) 79–82.

validate the algorithm.[74] Image recognition illustrates this point. If the training data does not contain images representing future uses, like images with different kinds of backgrounds, this can lead to recognition errors.[75] Bias does not only result from underrepresentation, where, for instance, image recognition training data contains fewer images of Black people or if training data for recruiting purposes includes few examples of successful female employees. Overrepresentation can also cause bias. 'Racial profiling', for example police stops targeting people of colour, typically lead to a much higher detection rate for people of colour than for the White population, which then suggests a – biased – statistical correlation between race and crime rate.[76]

Several factors might lead to the use of unrepresentative datasets. Representative datasets are often unavailable in contemporary societies shaped by inequalities. Moreover, existing datasets might be outdated,[77] designers might simply not realise that data is unrepresentative, or designers might be influenced by stereotypes or discriminatory preferences. If statistical assumptions cannot be properly reassessed, this might also lead to unrepresentative data, like when predictions concerning creditworthiness can only be verified with regard to the credits granted (not the credits that were denied) or if predictions concerning recidivism can only be controlled with regard to the decisions granting parole (not the decisions refusing parole).

b. Labelling Bias

Labelling, or the attribution of characteristics influenced by stereotypes or discriminatory preferences, can also induce bias.[78] Data not only refers to objective facts (e.g. the punctual discharge of financial obligations, high sales results), but also to subjective assessments (e.g. made on an evaluation platform or in job references). As a consequence, target variables (step 2), but also training and validation data (steps 1 and 4) and the predictor variables used in the predicting model (step 3), can relate either to objective facts or to subjective assessments. These assessments may reflect discriminatory prejudices and stereotypes as was shown for legal exams[79] or the evaluation of teachers.[80] In addition to that, discriminatory assessments might also result in – biased – facts, for example if the police stops or arrests members of minority groups at a disproportionately high level.

c. Feature Selection Bias

Feature selection bias means that relevant characteristics are not sufficiently taken into account.[81] Algorithms consider all data available when establishing correlations used for predictions (steps 1, 2, 4). Car insurance companies, for example, traditionally rely on specific data

[74] Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 46.

[75] *Cf.* the recognition of wolves and huskies M Ribeiro, S Singh, and C Guestrin, 'Why Should I Trust You?' in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) 1135 (1142).

[76] F Schauer, *Profiles, Probabilities, and Stereotypes* (2003) 194; B Harcourt, *Against Prediction. Profiling, Policing and Punishing in an Actuarial Age* (2007) 145 (hereafter Harcourt, *Against Prediction*).

[77] Kleinberg and others, *Discrimination in the Age of Algorithms* (n 6), 41 ('zombie predictions').

[78] Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 50–51; Barocas and Selbst, 'Big Data's Disparate Impact' (n 8) 681; Orwat, *Diskriminierungsrisiken* (n 8) 77–78.

[79] Female and immigrant students receive lower grades E Towfigh, C Traxler, and A Glöckner, 'Geschlechts- und Herkunftseffekte bei der Benotung juristischer Staatsprüfungen' (2018) 5 *Zeitschrift für Didaktik der Rechtswissenschaften* 115.

[80] A Özgümüs and others, 'Gender Bias in the Evaluation of Teaching Materials' (2020) 11 *Frontiers in Psychology* 1074.

[81] *Cf.* Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 52–53; Barocas and Selbst, Big Data's Disparate Impact' (n 8) 688.

concerning the vehicle (car type, engine power) and the driver(s) (age, address, driving experience, crash history; in the past also gender[82]) to specify the risk of a traffic accident. One can assume, however, that other types of data like an aggressive or defensive driving style correlate much stronger with the risk of accident than age (or gender).[83] Instead of imposing particularly high insurance premiums upon young (male) novice drivers, insurance companies could define categories of premiums according to the driving style and thus avoid discrimination based on age (or gender). Similarly, assessing the credit default risk could be based on meaningful features like income and consumer behaviour instead of relying on the borrower's address, which disadvantages the residents of poorer quarters ('redlining').[84]

d. Error Rates

Finally, statistical predictions also generate errors. Therefore, one has to accept certain error rates, such as false positives (e.g. predicting a high risk of recidivism where the offender does not reoffend) and false negatives (predicting a low risk of recidivism where the offender actually reoffends). It is now a matter of normative assessment which error rates seem acceptable for which kinds of decisions, for example for denying a credit or adding someone to the no-fly list. Moreover, when defining the target of profiling (step 2), the designers of algorithms must also decide how to allocate different error rates among different societal groups. If the relevant risks are not distributed evenly among different societal groups (say, if women have a higher risk of being genetic carriers of a disease than men or if men have a higher risk of recidivism than women), it is mathematically impossible to allocate similar error rates to all the affected groups, either overall for women and men, or for women and men within the group of false negatives or false positives respectively.[85] This problem was first detected and discussed in the context of predicted recidivism, where differing error rates manifested for Black versus White criminal offenders.[86] It follows from the trade-off that algorithms' designers can influence the allocation of error rates, and that regulators could shape this decision through legal rules.

IV. JUSTIFYING DIRECT AND INDIRECT FORMS OF DISCRIMINATORY AI: NORMATIVE AND TECHNOLOGICAL STANDARDS

The previous section highlighted different causes for discrimination in decision-making based on profiling. This section now turns to the question of justification, and argues that these causes are a relevant factor for the proportionality of direct or indirect discrimination. After specifying the proportionality framework (1), this section develops general considerations concerning statistical discrimination or group profiling (2) and examines the methodology of automated profiling (3) before turning to the difference between direct and indirect discrimination (4).

[82] This practice has been banned by the CJEU, Case C-236/09 *Test-Achats* [2011].

[83] On this example *cf.* Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 52–53.

[84] Barocas and Selbst, 'Big Data's Disparate Impact' (n 8) 689.

[85] J Kleinberg, S Mullainathan, and M Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' in C Papadimitrou (ed), *8th Innovations in Theoretical Computer Science Conference* (ITCS 2017) 43:1 (hereafter Kleinberg, Mullainathan, and Raghavan, 'Inherent Trade-Offs'); K Zweig and T Krafft, 'Fairness und Qualität Algorithmischer Entscheidungen' in M Kar, B Thapa, and P Parycek (eds), *(Un)Berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (2018) 204 (213–218) (hereafter Zweig and Krafft, 'Fairness und Qualität').

[86] Critically Angwin and others, 'Machine Bias' (n 4); on the problem Kleinberg, Mullainathan, and Raghavan, 'Inherent Trade-Offs' (n 85); Zweig and Krafft, 'Fairness und Qualität' (n 86); Cofone, 'Algorithmic Discrimination' (n 10) 1433–1436.

1. Proportionality Framework

The justification of discriminatory measures regularly includes proportionality.[87] EU law, for example, speaks of 'appropriate and necessary' means[88] of 'proportionate' genuine and determining occupational requirements[89] or, in the general limitation clause of Article 52 (1) Charter of Fundamental Rights, of 'the principle of proportionality'. Different legal systems vary in how they define and assess proportionality. The European Court of Human Rights applies an open 'balancing' test with respect to Article 14 ECHR,[90] and the European Court of Justice normally proceeds in two steps, analysing the suitability (appropriateness) and the necessity of the measure at stake.[91] In German constitutional law and elsewhere,[92] a three-step test has been established. According to this test, proportionality means that a (discriminatory) measure is suitable to achieve a legitimate aim (step 1), necessary to achieve this aim, meaning that the aim cannot be achieved by less onerous means (step 2), and appropriate in the specific case, where the legal interest pursued by a discriminatory measure outweighs the conflicting legal interest of non-discrimination (step 3). This three-step test will be used as an analytical tool to flesh out arguments that are relevant for justifying differential treatment or detrimental effect as a result of profiling and decision-making. Before this analysis, some aspects merit clarification.

a. Proportionality as a Standard for Equality and Anti-Discrimination

Some legal scholars claim that the notion of proportionality is only useful for assessing the violation of freedoms, not of equality rights. According to this view, an interference with a freedom, such as limits on the freedom of speech, constitutes a harm that needs to be justified with respect to a conflicting interest, such as protection of minors. In contrast, unequal treatment is omnipresent. It does not constitute prima facie harm (e.g. different laws for press and media platforms), and it typically does not pursue conflicting objectives. Rather, it reflects existing differences. To illustrate, different rules on youth protection for the press and for media platforms are not necessarily in conflict with youth protection. Rather, they result from different risks emanating from the press and media platforms.[93] Thus, in order to justify differential treatment one has to show that this differentiation follows 'acceptable standards of justice'

[87] On justification norms *cf.* n 46–48.

[88] E.g. with respect to direct discrimination Article 4(5) Goods and Services Sex Discrimination Directive 2004/113/EC; with respect to indirect discrimination e.g. Article 2(2)(b)(i) Framework Directive 2000/78/EC; Article 2(2)(b) Race Directive 2000/43/EC; Article 2(b) Goods and Services Sex Discrimination Directive 2004/113/EC; Article 2(1)(b) Equal Treatment Directive 2006/54/EC.

[89] E.g. with respect to direct discrimination Article 4(1) Framework Directive 2000/78/EC; Article 4 Race Directive 2000/43/EC; Article 14(2) Equal Treatment Directive 2006/54/EC.

[90] Harris and others, *European Convention on Human Rights* (n 48) 774; B Rainey and others, *The European Convention on Human Rights* (7th ed. 2017) 646–647.

[91] T Tridimas, 'The Principle of Proportionality' in R Schütze and T Tridimas (eds), *Oxford Principles of European Union Law*, Vol 1, 243, 247 (hereafter Tridimas, 'The Principle of Proportionality'); see, for example, CJEU, Case C-555/07 *Kücükdeveci* [2010] para 37–41; CJEU, Case C-528/13 *Léger* [2015] para 58–68; CJEU, Case C-157/15 *Achbita* [2017] para 40–43; CJEU, Case C-914/19 *GN* [2021] para 41–50; but note also the three-prong test including proportionality in the narrower sense, for example, in CJEU, Case C-83/14 *CHEZ* [2015] para 123–127.

[92] R Poscher in M Herdegen and others (eds), *Handbook on Constitutional Law* (2021) § 3 (forthcoming) (hereafter Poscher in 'Handbook on Constitutional Law'); N Petersen, *Verhältnismäßigkeit als Rationalitätskontrolle* (2015) (hereafter Petersen, *Verhältnismäßigkeit*); on the spread of this concept A Stone Sweet and J Mathews, 'Proportionality Balancing and Global Constitutionalism' (2008) 47 *Columbia Journal of Transnational Law* 72.

[93] The example is mine. The proportionality test is criticised by U Kischel, 'Art. 3 GG' in V Epping and C Hillgruber (eds), *BeckOK Grundgesetz* (47th ed. 2021) para 34–38a (hereafter Kischel, 'Art. 3 GG'), with further references.

reflecting 'relevant' differences,[94] or that the objective reasons outweigh the inequality impairment.[95] Only if differential treatment is meant to promote an 'external' objective unrelated to existing differences[96] should a proportionality assessment be made, according to some scholars.[97]

Nevertheless, the proportionality framework remains useful for the task of justifying discriminatory AI. The aforementioned proportionality scepticism seems partly motivated by the concern that equality rights and justification requirements must not expand uncontrollably. However, this valid point only applies to general equality rights in the context of which this concern was voiced, not to anti-discrimination law. Favouring men over women and *vice versa* does constitute *prima facie* harm, and justifying this differential treatment requires strict scrutiny and the consideration of less harmful alternative measures. In part, proportionality seems to be rejected as a justification standard because its criteria are too unclear. However, the proportionality assessment is flexible enough to take into account the characteristics of discriminatory measures. Thus, the proportionality test should evaluate whether using a particular differentiation criterion (like gender) is suitable, necessary, and appropriate for reaching the differentiation aim (e.g. setting appropriate insurance premiums, stopping tax evasion). For differential treatment based on profiling, this indeed implies that the differentiation criterion and the differentiation aim are not in conflict with each other as the decision-making responds to the different risks predicted as a result of profiling. A proportionality assessment now allows for strict scrutiny of both decision-making and profiling. This advantage of the proportionality test becomes increasingly important as profiling replaces older methods of differentiating between people. Moreover, a second advantage of the proportionality approach is its dual use for both direct and indirect discrimination. The detrimental effect of a facially neutral measure must not be justified with reference to existing differences. Quite the contrary, it must be justified with reference to an 'external' objective and proportionate means to achieve this objective.[98] Thus, apart from the fact that the law calls for proportionality, there are good reasons to stick to this standard, particularly for an assessment of profiling.

b. Three Steps: Suitability, Necessity, Appropriateness

In a nutshell, the proportionality test entails three simple questions: first, do the measures work, that is, does profiling and decision-making promote the (legitimate) aim (suitability)? Second, are there alternative, less onerous means of profiling and decision-making to achieve this aim (necessity)? Third, is the harm caused by profiling and decision-making outweighed by other interests (appropriateness)? If questions one and three can be answered in the affirmative and if question two can be answered negatively, the measure is proportionate and justified.

Note that this counting method does not include the preceding step of verifying that a measure pursues a legitimate aim, nor does it comprise the rarer consideration that the means

[94] S Huster, 'Art. 3' in KH Friauf and W Höfling (eds), *Berliner Kommentar zum Grundgesetz* (50th supplement 2016) para 89 (hereafter Huster, 'Art. 3').

[95] Kischel, 'Art. 3 GG' (n 93) para 37.

[96] S Huster, 'Gleichheit und Verhältnismäßigkeit' (1994) 49 *Juristenzeitung* 541, 543 (hereafter Huster, 'Gleichheit und Verhältnismäßigkeit') gives the examples of (1) different taxation based on different income which he qualifies as reflecting existing inequalities ('internal objective') and (2) different taxation aimed at stimulating the construction industry, providing tax relief for builders, which he qualifies as 'external objective'.

[97] Huster, 'Gleichheit und Verhältnismäßigkeit' (n 96) 549; Huster, 'Art. 3' (n 94) para 75–86, with further references.

[98] One can draw a parallel between direct and indirect discrimination on the one hand and *Huster's* idea of 'internal' and 'external' objectives in equality cases on the other hand (n 94 and 96).

used for pursuing this aim is itself legitimate.[99] It can be assumed that the aims pursued by decision-making based on profiling pursue legitimate aims, such as finding and hiring the most qualified applicant or monitor persons inclined to commit a crime. This article will also neglect the possibility that the means itself is prohibited. Profiling might be prohibited per se, for example, if past human actions are assessed individually. An individual criminal conviction or student performance grade cannot be based on statistical predictions concerning recidivism among certain groups of offenders or based on certain schools' performance.[100]

Turning to the 3-step test, it should be emphasised that it refers to profiling and decision-making, this means to two interrelated, but different acts. It is the decision that needs to be justified under non-discrimination law for involving different treatment or for causing detrimental effect. However, as far as this decision is based on a prediction resulting from profiling, profiling as an instrument of prediction must also be proportionate. Profiling is proportionate if it generates valid predictions (suitability, step 1), if alternative profiling methods that generate equally good predictions at lower costs do not exist (necessity, step 2), and if the harm of profiling is outweighed by its benefits (appropriateness, step 3). In addition, other aspects of the discriminatory decision also come under scrutiny, notably the harm of a decision (for example a police control involves a different sort of harm than a flight ban).[101]

Some proportionality scholars doubt that steps 2 and 3 can be meaningfully separated.[102] The European Court of Justice (ECJ), which typically applies a 2-step test comprising suitability and necessity, sometimes includes elements of balancing in its reasoning at the second step,[103] but increasingly also resorts to the 3-step test.[104] This chapter submits that it is helpful to separate steps 2 and 3. In step 2, the measure in question is compared to alternative measures which are equally effective in achieving a particular aim, for example, different profiling methods equally good at predicting a risk. If an alternative means generates more costs or curtails other rights, the conditions 'equally suitable' and 'less burdensome' are not met.[105] This means comparing both normative and factual burdens for different groups of people: the persons affected by the measure under review, third parties that might be affected by alternative measures, and the decision-maker. An alternative profiling method, for example, could place a different burden on the persons affected by the measure under review (e.g. by using more personal data and thus limiting privacy). An alternative profiling method could also place a burden on third parties (e.g. if the alternative method yields negative profiling results followed by disadvantageous decisions). Finally, an alternative profiling method could also burden the decision-maker because the method requires more resources such as time or money. These considerations involve value

[99] *Cf.* Poscher in 'Handbook on Constitutional Law' (n 92).

[100] In the UK, it was planned to use an A-level algorithm predicting grades in 2020 as the A-level exams were cancelled due to COVID-19. The algorithm was meant to take into account the teachers' assessment of individual pupils and the performance of the respective school in past A-level exams in order to combat inflation in grades. The algorithm would have had disadvantaged good pupils from state-run schools with ethnic minorities. The project was cancelled after public protest. *Cf.* Wachter, Mittelstadt, and Russell, 'Bias Preservation' (n 12) 1–6.

[101] On these points *cf.* Sub-sections IV 2 and 3.

[102] Moreover, it is disputed that rational criteria exist for the balancing exercise of step 3. *Cf.* T Kingreen and R Poscher, *Grundrechte Staatsrecht II* (36th ed. 2020) § 6 para 340–347; for an in-depth analysis on the criticism of balancing and its underlying, see N Petersen, 'How to Compare the Length of Lines to the Weight of Stones: Balancing and the Resolution of Value Conflicts in Constitutional Law' (2013) 14 *German Law Journal* 1387.

[103] Tridimas, 'The Principle of Proportionality' (n 91); *cf.* also G de Burca, 'The Principle of Proportionality and Its Application in EC Law' (1993) 13 *Yearbook of European Law* 105, 113–114.

[104] B Oreschnik, *Verhältnismäßigkeit und Kontrolldichte* (2018) 158–178, 219–227.

[105] Poscher in 'Handbook on Constitutional Law' (n 92) paras 63–67.

assessments, as different burdens have to be identified and weighed. It is not surprising that some legal systems prefer to see these considerations as part of the balancing test (step 3), whereas other legal systems address reasonable alternative measures under the heading of necessity only (step 2).[106] It is nevertheless a useful analytical tool to distinguish between less onerous alternative means (step 2) and other alternative means (step 3).

Finally, it should be emphasised that by treating proportionality as a general issue, this article does not mean to downplay the particularities of specific justification provisions or to conceal the different harms caused by different forms of discrimination. Particularly severe forms of direct discrimination will hardly be justifiable at all (like direct discrimination on grounds of race) or merit very strict scrutiny (for example direct discrimination on grounds of gender which can be justified based on biological differences), other forms might be much easier to justify depending on the circumstances. Furthermore, a distinction must also be drawn between decisions made by the state and by private actors. Even if anti-discrimination law covers both, the state is directly bound by fundamental rights including equality and non-discrimination. By contrast, the choices and actions of private actors are protected by fundamental freedoms such as freedom of contract or freedom to conduct a business, leading to a stricter burden of justification for state actors than for private actors. The point of this article is to elaborate on the commonalities of discriminatory decision-making based on profiling, and to show the relevant aspects for assessing its legality.

2. *General Considerations Concerning Statistical Discrimination/Group Profiling*

In the context of discriminatory profiling and decision-making, it is useful to distinguish general aspects of proportionality that are known from non-automated forms of statistical discrimination (this section), and specific aspects of automated group profiling (IV.3.). Note that the terms 'statistical discrimination' and decision-making based on 'group profiling' designate the same phenomena.[107] The first term is long-established, while the term 'group profiling' is mainly used in the context of automated profiling. Both refer to differential treatment or detrimental effect that results from statistical predictions and affects groups defined by sensitive characteristics or its members. Before looking at specific issues of the methodology of profiling in the next section, this section will highlight some arguments relevant for the proportionality test.

a. Different Harms: Decision Harm, Error Harm, Attribution Harm

As a starting point, one can distinguish different harms stemming from profiling and decision-making.[108] The decision itself contains negative consequences corresponding to a varying degree of 'decision harm': a denial of goods (no credit), bad contract terms (high insurance premiums), a denial of chances (no job interview), or investigations (a police control). 'Decision harms' arise in human and automated decisions alike. But some forms of 'decision harm' are typical of decisions based on profiling. Profiling is meant to overcome an information deficit (Who is a qualified employee? Which person is about to commit a crime?). Therefore, many decisions tend to be part of an information gathering process: Some job applicants are chosen for a job interview, while others are refused right away. Some taxpayers are singled out for an audit, while other filers' tax declarations are accepted without further review. It is important to recognise that

[106] Petersen, *Verhältnismäßigkeit* (n 92), 144–147, 258–262, for example, argues comprehensively that it might be easier for well-established, powerful courts to openly apply a balancing test than for other courts.

[107] See Sub-section III 1.

[108] See also Britz, *Einzelfallgerechtigkeit* (n 61) 120–136, albeit with different classifications.

these decisions involve a harm of their own. They attribute opportunities and risks which can be very relevant for the individual person, but they can also lead to the deepening of existing stereotypes and inequalities.

Other harms relate to profiling. Statistical predictions generated by profiling have a certain error rate, which means that false positives (like honest taxpayers flagged for the risk of fraud) or false negatives (as creditworthy consumers with a low credit score) suffer from the negative consequences of a decision. This sort of 'error harm' is already known as 'generalisation harm' in jurisprudence. Legal systems are based on legal rules which, by definition, apply in a general manner, as opposed to decisions based on specific issues targeting specific individuals. A general rule will often be overinclusive. For example an age limit for pilots addresses pilots' statistically decreasing flying ability with age, but it also applies to persons who are still perfectly fit to fly.[109] This sort of 'generalisation harm' can be quantified in the process of automated profiling as error rates. Finally, group profiles also carry the risk of 'attribution harm' if they associate all members of a group with a negative characteristic, e.g. Black people with higher criminality or women with lower performance. The degree of 'attribution harm' can also vary: some characteristics predicted by profiling can be embarrassing or humiliating (like crime, low work performance, confidential health data), while others are not problematic (e.g. high purchasing power). Some of these negative attributions are visible to others (such as police disproportionately stopping or searching Black people), while others remain hidden in the algorithm. Some attributions confirm and reinforce existing stereotypes, while others run counter to existing prejudices (for example a good driving record for women). Some attributions can be corrected in the individual case (e.g. if a police check does not yield a result), while others remain unrefuted.

Under the proportionality test, these harms, the varying degrees of harm evoked in particular instances, are relevant for steps 2 and 3, that is, for assessing whether alternative means are less onerous (evoke less harm) than the measure at hand (necessity, step 2), and for balancing the conflicting interests (appropriateness, step 3).

b. Alternative Means: Profiling Granularity and Information Gathering

After defining the distinct harms of profiling and decision-making, we can now turn to concrete strategies to better reconcile conflicting interests. This is again either a matter of necessity (step 2) or appropriateness (step 3). The measure at issue is not necessary if an alternative means is equally suitable to reach a particular aim without imposing the same burden, and the measure is not appropriate if it is reasonable to resort to an alternative measure that better reconciles the conflicting interests.

This chapter outlines two possible alternative means for decisions based on profiling. The first concerns the granularity of the profiles. Sophisticated profiles obtained from a wealth of data are more accurate than simple profiles based on a few data points only. If decisions are based on simple profiles, then the above-mentioned 'generalisation harm' can result from both profiling and decision-making, as larger groups of people count among the false positives and false negatives[110] and larger groups also suffer the negative effect of a decision. Blood donation, for example, should not lead to the transmission of HIV. In order to reduce this risk, one could exclude several groups from blood donation: homosexuals, male homosexuals, only sexually active male homosexuals, or only sexually active male homosexuals engaging in behaviour

[109] *Cf.* CJEU, Case C-190/16 *Fries* [2017].
[110] On error rates see also Sub-section III 2(d).

which puts them at a high risk of acquiring HIV. The more the group is defined, the smaller the number of people affected by a prohibition of blood donation.[111] As a consequence, the higher accuracy of fine-granular group profiles must, therefore, be weighed against the advantages of simple group profiles such as data minimisation or simplicity. The need for granular profiles is expressed, for example, in the German implementation of the European Passenger Name Record (PNR) system. The EU PNR Directive provides that air passengers are assessed with respect to possible involvement in terrorism or other serious crime. This is done by comparing passenger data against relevant databases and pre-determined criteria (i.e. by profiling), and these criteria need to be 'targeted, proportionate and specific'.[112] The German Air Passenger Data Act implementing this provision stipulates that the relevant features (i.e. factors providing ground for suspicion, as well as exonerating factors) must be combined 'such that the number of persons matching the pattern is as small as possible'.[113]

Second, as profiling helps address information deficits, alternative means of coping with these deficits can also be a relevant aspect of the proportionality test. If information is particularly important, fully clarifying the facts can be preferable to profiling, provided that this is feasible and that the resources are available. Take the example of airport security screening. Screening of air passengers and their luggage items is not confined to a certain sample of 'high risk' passengers but extends to all passengers. Regarding the blood donation example, systematically screening all blood donations for HIV could be an alternative means to refusing sexually active male homosexuals to donate blood.[114] Similar forms of full fact-finding are also conceivable in the context of automation, although they create costs and they entail the large-scale processing of personal data. Another method of reconciling the need for information and non-discrimination is randomisation, this means gathering information at random. If only a fraction of tax returns can be scrutinised by the fiscal authorities, these tax returns can be chosen at random or based on the profile of a tax evader. Using risk profiles might seem to allocate resources more efficiently, but randomisation has other advantages: it burdens all taxpayers equally and prevents discriminatory effects.[115] In addition, it might also be more efficient and less susceptible to manipulation because taxpayers cannot game the algorithm.[116]

3. *Methodology of Automated Profiling: A Right to Reasonable Inferences*

This section turns to the methodology of automated profiling, which has a decisive impact on the possible harms of discriminatory AI.[117] It looks at legal sources for explicit and implicit methodology standards and links them to the elements of the proportionality test. As a result, this section claims that a 'right to reasonable inferences'[118] already exists in the context of discriminatory AI.

[111] CJEU, Case C-528/13 *Léger* [2015] para 67.
[112] Article 6(4) Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime.
[113] Section 4(3) Passenger Name Record Act of 6 June 2017 *Bundesgesetzblatt* I 1484, as amended by Article 2 of the Act of 6 June 2017 *Bundesgesetzblatt* I 1484.
[114] CJEU, Case C-528/13 *Léger* [2015] para 64.
[115] Harcourt, *Against Prediction* (n 76) 237.
[116] The German automated risk management system which selects tax returns for human review is complemented by randomised human tax reviews, Section 88(5)(1) German Fiscal Code of 1 October 2002 *Bundesgesetzblatt* I 3866, last amended by Article 17 of the Act of 17 July 2017 *Bundesgesetzblatt* I 2541.
[117] See Sub-section III 2.
[118] Called for by Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

a. Explicit and Implicit Methodology Standards

As opposed to other activities, such as operating a nuclear power plant or selling pharmaceuticals, developing and using profiling algorithms does not require a permission issued by a state agency. Operators of nuclear power plants in Germany, for example, must show that 'necessary precautions have been taken in accordance with the state of the art in science and technology against damage caused by the construction and operation of the installation' before obtaining a licence,[119] and pharmaceutical companies need to prove that pharmaceuticals have been sufficiently tested and possess therapeutic efficacy 'in accordance with the confirmed state of scientific knowledge'[120] before obtaining the necessary marketing authorisation. The referral to the 'state of the art in sciences and technology' or the 'confirmed state of scientific knowledge' implies that methodology standards developed outside the law, for example in safety engineering or pharmaceutics, are incorporated into the law. Currently, there is no similar *ex ante* control of profiling algorithms, which means that algorithms are not measured against any methodological standards in order to qualify for a permission. This situation might change, of course. The German Data Ethics Commission, for example, suggests that algorithmic systems with regular or serious potential for harm should be covered by a licensing procedure or preliminary checks.[121]

But the lack of a licensing procedure does not mean that methodology standards for algorithmic profiling do not exist. Some legal norms explicitly refer to methodology, and implicit methodological standards can also be found in the general justification test for discrimination. These standards may be enforced – *ex post* – by affected individuals who bring civil or administrative proceedings, or by public agencies like data protection authorities or anti-discrimination bodies who control actors and fine offenders.[122]

Legal norms that explicitly state methodology requirements for profiling and decision-making exist. The German Federal Data Protection Act, for example, regulates some aspects of scoring, such as the use of a probability value for certain future action by a natural person and, hence, a particular form of profiling. The statute stipulates that 'the data used to calculate the probability value are demonstrably essential for calculating the probability of the action on the basis of a scientifically recognised mathematic-statistical procedure'.[123] Similar requirements can be found in insurance law. The Goods and Services Sex Discrimination ('Unisex') Directive 2004/113/EC contains an optional clause enabling states to permit the use of sex as a factor in insurance premium calculation and benefits 'where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data'.[124] After the ECJ declared this clause invalid due to sex discrimination,[125] the methodology requirement remains nevertheless relevant for old insurance contracts and provides an inspiration for national standards such as the German General Act on Equal Treatment. This statute, which implements EU anti-

[119] Section 7(2)(3) German Atomic Energy Act of 15 July 1985 *Bundesgesetzblatt* I 1565, as last amended by Article 3 of the Act of 20 May 2021 *Bundesgesetzblatt* I 1194.

[120] Section 25(2)(2 and 4) German Medicinal Products Act of 12 December 2005 Bundesgesetzblatt I 3394, as last amended by Article 11 of the Act of 6 May 2019 *Bundesgesetzblatt* I 646. Emphasis by author.

[121] German Data Ethics Commission, *Opinion of the Data Ethics Commission* (2019) 195 (hereafter German Data Ethics Commission, *Opinion*).

[122] *Cf.* the broad powers of the data protection authorities under Articles 58, 70, 83–84 GDPR.

[123] Section 31(1)(2) German Federal Data Protection Act of 30 June 2017 *Bundesgesetzblatt* I 2097, as last amended by Article 12 of the Act of 20 November 2019 *Bundesgesetzblatt* I 1626; a similar provision can also be found in Section 10 (2)(1) Bancing Act (Kreditwesengesetz). Note that it is disputed whether Section 31 Federal Data Protection Act is in conformity with the GDPR, (i.e. whether it is covered by one of its opening clauses).

[124] Article 5(2) Unisex Directive 2004/113/EC.

[125] CJEU, Case C-236/09 *Test-Achats* [2011].

discrimination law and establishes additional national standards of anti-discrimination law, also contains a methodology requirement for calculating insurance premiums and benefits: 'Differences of treatment on the ground of religion, disability, age or sexual orientation [...] shall be permissible only where these are based on recognised principles of risk-adequate calculations, in particular on an assessment of risk based on actuarial calculations which are in turn based on statistical surveys.'[126] Note that these rules refer to recognised procedures of other disciplines like mathematics, statistics, and actuarial sciences which guarantee that certain aspects of profiling are reasonable from a methodological point of view, that is, that using personal data is 'essential' for probability calculation or that relying on a protected characteristic like sex is a 'determining factor' for risk assessment.

In other contexts, statutes do not refer to methodology in the narrower sense, but to other aspects related to the validity of profiling and establish review obligations. Thus, the EU PNR Directive stipulates that the profiling criteria have to be 'regularly reviewed'.[127] The risk management system used by the German revenue authorities must ensure that 'regular reviews are conducted to determine whether risk management systems are fulfilling their objectives'.[128]

But even if explicit standards do not exist, implicit methodological requirements flow from the justification test – in other words, the proportionality test – of anti-discrimination law. Discriminatory decisions based on automated profiling need to pass the proportionality test, and this includes the methodology of profiling.[129] It is a matter of suitability (step 1) that automated profiling produces valid probability statements. Only then does it further a legitimate goal if a discriminatory decision is based on the result of profiling. Furthermore, it needs to be discussed in the context of necessity (step 2) and appropriateness (step 3) whether a different methodology of profiling and decision-making would have a less discriminatory effect. If the profiling methodology can be improved, if its harms can be reduced, the costs and benefits of these improvements will be relevant for considerations of necessity and appropriateness.

For the sake of completeness, this chapter argues that methodological profiling standards can also be derived from data protection law. In accordance with Article 6(1) of the GDPR the processing of personal data, which is essential for profiling a particular person,[130] requires a legal basis. All legal bases for data processing except consent demand that data processing is 'necessary' for certain purposes, that is, for the performance of a contract,[131] for compliance with a legal obligation,[132] for the performance of a task carried out in the public interest,[133] or for the purposes of legitimate interests.[134] For automated profiling and decision-making, Article 22(2) and (3) GDPR also require suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, which includes non-discrimination. Thus, the necessity test of Article 6 (1) GDPR and the safeguarding clause of Article 22(2) and (3) GDPR also imply a minimum standard of profiling methodology. Data processing for profiling is only necessary for the

[126] Section 20(2) German General Act on Equal Treatment of 14 August 2006 *Bundesgesetzblatt* I 1897, as last amended by Article 8 of the SEPA Accompanying Act of 3 April 2013 *Bundesgesetzblatt* I 610. *Cf.* Section 33(5) General Act on Equal Treatment, on old insurance contracts and gender discrimination.

[127] Article 6(4) PNR Directive (EU) 2016/681.

[128] Section 88(5) German Fiscal Code of 1 October 2002 *Bundesgesetzblatt* I 3866; 2003 I 61, last amended by Article 17 of the Act of 17 July 2017 *Bundesgesetzblatt* I 2541.

[129] See Sub-section IV 1(b).

[130] This is the third step of the profiling process, see II1.

[131] GDPR, Article 6(1)(b).

[132] GDPR, Article 6(1)(c).

[133] GDPR, Article 6(1)(e).

[134] GDPR, Article 6(1)(f).

above-mentioned goals, if the profiling method produces valid predictions and if no alternative profiling method exists which makes equally good predictions while discriminating less. Similar standards can be derived from Article 22 GDPR for automated decision-making based on profiling.

These implicit methodological standards can be developed from the proportionality requirements of anti-discrimination and data protection law even if the legislator has also enacted specific methodological standards with a limited scope of application. Specific methodological standards have long existed in areas of law like insurance and credit law, which refer to established mathematical-statistical standards. Anti-discrimination lawyers, however, have only recently started to call for methodological standards of profiling,[135] long after today's anti-discrimination laws were formulated.[136] Admittedly, the 2016 GDPR addresses the dangers of profiling without also formulating an explicit legal methodological requirement. But Recital 71 requires that 'the controller should use appropriate mathematical or statistical procedures for the profiling [...] in a manner [...] that prevents [...] discriminatory effects'.[137] This non-binding recital expresses the lawmakers' intentions and can help to interpret the legal obligations of the GDPR. Several provisions of GDPR and other recitals also show that the Regulation intends to effectively address the dangers of profiling, including the danger of discrimination.[138] As a consequence, even if the GDPR does not establish an explicit profiling methodology, a minimum standard is implicitly included in the requirement of 'necessary' data protection. In this respect, profiling differs from activities governed by standards outside of data protection law. For example, evaluating exam papers and inferring from these pieces of personal data whether the candidate qualifies for a certain grade follows criteria that have been developed in the examination subject. These criteria cannot be found in data protection law.[139] Inferring information by means of profiling, however, is an activity inextricably linked to data processing and clearly covered by the GDPR.

This minimum standard of a proportionate profiling methodology does not amount to a free-standing 'right to reasonable inferences'[140]. It is a justification requirement triggered by discrimination, this means by different treatment and detrimental impact. However, many decisions based on profiling will involve different treatment or detrimental impact. As a consequence, this

[135] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10) (2019).

[136] Article 21 European Charter of Fundamental Rights (2000), Framework Directive 2000/78/EC, Race Directive 2000/43/EC, Goods and Services Sex Discrimination Directive 2004/113/EC, Equal Treatment Directive 2006/54/EC; German General Equal Treatment Act (2006); not to mention Article 3 German Basic Law (1949) or Article 14 European Convention of Human Rights (1950).

[137] The full sentence reads: "In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect."

[138] Automated decision-making based on profiling is not only addressed in Article 22 GDPR, but also in Articles 13(2)(f), 14(2)(g), 15(1)(h) GDPR (rights to information), Article 35(3)(a) GDPR (data protection impact assessment), Article 47(2)(e) GDPR (binding corporate rules), Article 70(1)(f) GDPR (guidelines of the European Data Protection Board); profiling as such is addressed in Article 21(1) and (2) GDPR (right to object to certain forms of profiling); moreover Recitals 24, 60, 63, 70–73, 91 concern aspects of profiling. The aim to prevent discrimination is not only expressed in Recital 71, but also in Recital 75 (concerning risks to the rights and freedoms resulting from data processing) and in Recital 85 (concerning damage due to personal data breach).

[139] This is why the right to rectification does not extend to incorrect answers, CJEU, *Nowak* C-434/16, [2017] (n 52–57); *cf.* already Sub-section II 2.

[140] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

minimum standard of proportionate profiling methodology has a wide scope of application. What's more, this standard does not only entail the need for 'reasonable' inferences. Proportionality comprises more than the validity of inferences, it also calls for the least discriminatory methodology that is possible or that can be reasonably expected of the decision-maker.

b. Technical and Legal Elements of Profiling Methodology

The practical challenge now lies in developing appropriate methodological standards.[141] From a technical point of view, disciplines such as data science, mathematics, and computer science shape these standards. At the same time, legal considerations play a decisive role as these methodological standards have a legal basis in the proportionality test. Both technical and legal elements are relevant for assessing the suitability (step 1), the necessity (step 2), and appropriateness (step 3) of profiling.

Returning to the elements of profiling[142] and to the factors identified as causing and affecting discriminatory decisions,[143] it is important to emphasise how technical *and* legal considerations are crucial in developing the right profiling methodology. In regards to error rates, first, it is a technical question to determine how reliable predictions are and how different error rates affect different groups of people depending on allocation decisions.[144] But it is a legal matter to define the minimum standard for the validity of profiling (relevant for suitability, step 1)[145] and to assess whether differences in error rates are significant when comparing the effects and costs of different profiling methods (relevant for necessity and appropriateness, steps 2 and 3). It is also a legal question whether different error rates among different groups are acceptable (i.e. necessary and appropriate).

Second, technical and legal assessments are also required for avoiding or evaluating bias, such as sampling, labelling, or feature selection biases, in the process of profiling. Sampling bias can be prevented by using representative training and testing data. How representative data sets can be obtained or created, and what amount of time, money, and effort this involves, are both technical questions. Moreover, data and computer scientists are also working on alternative methods to simulate representativeness by using synthetic data or processed data sets.[146] The legal evaluation includes the extent to which these additional efforts can be reasonably expected of the decision-maker. Similarly, there are attempts to counteract labelling bias by technical means, such as neutralising pejorative terms in target or predictor variables. But again, these options must also be assessed from a legal point of view, accounting for possible costs and legal harms, such as a loss of free speech in evaluation schemes. Feature selection bias can be reduced by replacing less relevant predictor variables with more relevant ones. Again, aspects of technical feasibility (for instance data availability) and technical performance (like error rate reduction)

[141] See also Orwat, *Diskriminierungsrisiken* (n 8) 114.
[142] Sub-section II 1.
[143] Sub-section III 2.
[144] See Sub-section III 2(d).
[145] Similar legal assessments can be found, for example, in Criminal Procedural Law regarding the reliability of DNA testing methods.
[146] Cofone, 'Algorithmic Discrimination' (n 10) 1431; German Data Ethics Commission, *Opinion* (n 121) 132. On further technical solutions see for example F Kamiran, T Calders, and M Pechenizkiy 'Techniques for Discrimination-Free Predictive Models' in T Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 223; S Hajian and J Domingo-Ferrer, 'Direct and Indirect Discrimination Prevention Methods' in T Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 241; S Verwer and T Calders, 'Introducing Positive Discrimination in Predictive Models' in T Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 255.

have to be combined with a legal assessment of technical and legal costs (e.g. a loss of data protection). These considerations concerning possible alternatives to avoid biases are part of the necessity and appropriateness test (steps 2 and 3). Apart from looking at error rates and bias, the proportionality assessment can finally also extend to the profiling model as such. One may argue, for example, that some decisions require a profiling model based on (presumed) causalities, not on mere correlations.

As a consequence, developing appropriate methodological profiling standards will require exchange and cooperation between lawyers and data and computer scientists. In this process, scientists have to explain the validity and the limits of existing methods as well as to explore less discriminatory alternatives, and lawyers have to specify and to weigh benefits and harms of these methods from a legal perspective.

4. *Direct and Indirect Discrimination*

One final aspect of justification concerns direct and indirect discrimination, or differential treatment and detrimental impact. Distinguishing direct and indirect discrimination has been a central tenet of discrimination law up to now. In the age of intelligent profiling, this distinction will become blurred, and indirect discrimination will become increasingly important.

a. Justifying Differential Treatment

In some contexts, even differential treatment based on protected characteristics such as gender, race, nationality, or religion is claimed to be justified based on statistical correlations. This is the case, for example, if unemployed women are less likely to get hired than men and job agencies allocate their services accordingly, if the Swedish minority in Finland has higher credit scores than the Finish majority and, hence, the Swedish can access credit more easily and at lower cost than the Finish, or if Muslims are presumed to have a stronger link to terrorism than the rest of the population and law enforcement agencies more closely scrutinise Muslims.[147] A justification of these forms of different treatment is not entirely ruled out. But the justification should be limited to extremely narrow conditions, especially in the case of particularly problematic characteristics. Even if race, gender, nationality, or religion happened to statistically correlate with certain risks, the harm inflicted by classifying people by these sensitive characteristics is too severe to be generally acceptable. It would not be appropriate (step 3), provided the measure passes the first two steps.[148]

b. Justifying Detrimental Impact

With regard to indirect discrimination, anti-discrimination law has to-date tended to concentrate on evident phenomena. In these cases, clear proxies exist, notably when employers disadvantage (predominantly female) part-time workers[149] or (predominantly Black) applicants who lack certain educational qualifications,[150] or when EU member states make rights or benefits conditional on domestic residence or language skills, which are requirements that are easily met by

[147] On these examples J Holl, G Kernbeiß, and M Wagner-Pinter, *Das AMS-Arbeitsmarktchancen-Modell* (2018) www.ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf; AlgorithmWatch, *Automating Society* (n 64) 59–60; Ungern-Sternberg, 'Religious Profiling'(n 65) 191–193.

[148] Ungern-Sternberg, 'Religious Profiling' (n 65) 205–211.

[149] CJEIU, C- 96/80 *Jenkins* [1981]; CJEU, C-170/84 *Bilka* [1986].

[150] *Griggs v. Duke Power Co*, 401 US 424 (1971).

most nationals, but not by EU foreigners.[151] Thus, indirectly disadvantaging women, Blacks, or aliens has to be justified by establishing that a measure is proportionate to reach a legitimate aim. However, do justification standards need to be equally high in the context of profiling, for example, if group profiles are much more refined and if overlaps with protected groups less clear? Or is it sufficient if profiling is based on a sound methodology? Lawyers will have to clarify why indirect discrimination is problematic and what amounts to such an instance of indirect discrimination.

There are good arguments in favour of extending stricter standards to situations in which proxies are less established and group profiles and protected groups overlap less significantly. Traditionally, one can distinguish 'weak' and 'strong' models of indirect discrimination.[152] According to the 'weak' model, indirect discrimination is meant to back the prohibition of direct discrimination by interdicting ways to circumvent direct discrimination.[153] 'Stronger' models pursue more far-reaching aims such as equality of chances[154] or equality of results correcting existing inequalities[155]. Furthermore, indirect discrimination might also be seen as a functional instrument to secure effective protection of non-discrimination where it overlaps with liberties like freedom of movement or freedom of religion.[156] Stronger models of indirect discrimination require that responsibilities and burdens of state and private actors are specified. In many cases it will be fair, for example, that employers do not have to bear the burden of existing societal inequalities, but that they refrain from perpetuating or deepening these inequalities.[157] Moreover, it seems helpful to specify particular harms caused in different situations that merit different forms of responses by non-discrimination law, for example redressing disadvantaging, addressing stereotypes, enhancing participation, or achieving structural change as proposed by *Sandra Fredman*.[158]

This chapter submits that the use of indirectly discriminatory algorithms also merits considerable scrutiny, for at least two reasons. First, big data analysis facilitates the linkage of innocuous data to sensitive characteristics. If internet platforms can infer characteristics like gender, sexual orientation, health conditions, or purchasing power from your online behaviour, they do not need to ask for this sensitive data in order to use it. This situation can be compared to the circumvention scenario that even 'weak' models of indirect discrimination intend to prevent. Second, it is increasingly difficult to distinguish between direct and indirect discrimination. The more complex profiling algorithms become and the more autonomously they operate, the more difficult it is to identify the relevant predictor variables (i.e. to tell whether profiling directly

[151] *Cf.* CJEU, C-152/73 *Sotgiu* [1974]; P Craig and J de Búrca, *EU Law* (7th ed., 2020) 796–797.

[152] Different weak and strong models are developed by Schiek, 'Indirect Discrimination'(n 41) 323–333 (circumvention vs. social engineering); Connolly, *Discrimination Law* (n 38) 153–156 (pretext, functional equivalency, quota model); Tobler, *Limits* (n 45) 24 (effectiveness of discrimination law and challenges the underlying causes of discrimination); see also Morris, 'On the Normative Foundations' (n 45) (corrective and distributive justice); M Grünberger, *Personale Gleichheit* (2013) 657–661 (hereafter Grünberger, *Personale Gleichheit*) (individual and group justice); S Fredman, 'Substantive Equality Revisited' (2016) 14 *I·CON* 713(hereafter Fredman '*Substantive Equality Revisited*') (formal and substantive equality); Wachter, Mittelstadt, and Russell, 'Bias Preservation'(n 12) para 2 (formal and substantive equality).

[153] This is a common position in Germany, *cf.* M Fehling, 'Mittelbare Diskriminierung und Artikel 3 (Abs. 3) GG' in D Heckmann, R Schenke, and G Sydow (eds) *Festschrift für Thomas Würtenberger* (2013) 668 (675).

[154] Wachter, Mittelstadt, and Russell, 'Bias Preservation'(n 12) para 2.1.1.

[155] Schiek, 'Indirect Discrimination' (n 41) 327.

[156] *Cf.* n 151 on freedom of movement; CJEU, Case C-157/15 *Achbita* [2017], and CJEU, Case C-188/15 *Bougnaoui* [2017] on freedom of religion, *cf.* also L Vickers, 'Indirect Discrimination and Individual Belief: Eweida v British Airways plc' (2009) 11 *Ecclesiastical Law Journal* 197.

[157] Grünberger, *Personale Gleichheit* (n 152) 660–661.

[158] Fredman, '*Substantive Equality Revisited*' (n 152).

includes a forbidden characteristic or not). In addition to this epistemic challenge, normative questions concerning the difference between direct and indirect discrimination arise. If a complex profile comprises 250 data points, among them one sensitive one (for instance gender) and 50 data points related to this sensitive characteristic (for example attributes typical of a certain gender), does using this profile involve different treatment or lead to detrimental impact? What if it cannot be established if the one sensitive data point was decisive for a particular outcome? The detrimental effect of profiling might be easier to prove than differential treatment because the output of profiling algorithms can be more easily tested than its internal decision-making criteria, especially with increasingly autonomous, self-learning, and opaque algorithms.[159] Because of this, it might be more helpful for the people affected and also more predictable for the users of profiling algorithms to assume indirect discrimination, but at the same time also to apply stricter scrutiny.

The broader the reach of indirect discrimination becomes, the more relevant the standards of justification will be.[160] Developing these standards will, therefore, be a crucial task in coping with discriminatory AI and in attributing responsibilities in the fight against factual discrimination. In part, these standards might be developed in view of existing ones. EU anti-discrimination law establishes, for example, that companies cannot justify discrimination against their employees by relying on customers' preferences, for these are not considered 'genuine and determining occupational requirements'.[161] The reasoning is also applicable to indirect forms of discrimination based on (predicted) customers' preferences and could therefore exclude a justification of policies or measures based on profiling. Moreover, as explained earlier, justification standards for both direct and indirect discrimination also depend on technical factors such as the possibilities and costs of avoiding discrimination. In the context of indirect discrimination, this might be relevant for errors in personalised (as opposed to group) profiling. Take the example of face recognition which yields particularly high error rates for Black women and low error rates for White men.[162] This could mean that Black women cannot use technical devices based on image recognition or that unnecessary law enforcement activities are directed against them. Provided that applying an algorithm with unequal error rates is covered by anti-discrimination law, that is, if it amounts to an apparently neutral practice that puts members of a protected group at a particular disadvantage,[163] one should ask how costly it would be to reduce error rates and how useful it would be to rely on other techniques until error rates are reduced.

V. CONCLUSION

Law is not silent on discriminatory AI. Existing rules of anti-discrimination law and data protection law do cover decision-making based on profiling. This chapter aims to show that the legal requirement to justify direct and indirect forms of discrimination implies that profiling

[159] On this F Pasquale, *The Black Box Society* (2015).

[160] Generally, on this point C McCrudden, 'The New Architecture of EU Equality Law after CHEZ' (2016) *European Equality Law Review* 1 (9).

[161] CJEU, C-188/15 *Bougnaoui* [2017] para 37–41.

[162] J Buolamwini and T Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 *Proceedings of Machine Learning Research* 1.

[163] The question which factual disadvantages are covered by anti-discrimination law cannot be treated here in detail. Traditionally, anti-discrimination law applies to differential treatment or detrimental impact as a result of legal acts (e.g. contractual terms, the refusal to conclude a contract, employers' instructions, statutes, law enforcement acts). But the wording of anti-discrimination law does not exclude factual disadvantages like a malfunctioning device, which might thus also trigger anti-discrimination provisions.

must follow methodological minimum standards. It remains a very important task for lawyers to specify these standards in case law or – preferably – legislation. For this, lawyers need to cooperate with data or computer scientists in order to assess the validity of profiling and to evaluate alternative methods by considering the discriminatory effects of sampling bias, labelling bias, and feature selection bias or the distribution of error rates.

The EU commission has recently published a proposal for the regulation of AI, the 'EU Artificial Intelligence Act'.[164] This piece of legislation would indeed specify relevant standards significantly. According to the proposal, AI systems classified as 'high risk' have to comply with requirements which reflect the idea that AI systems should produce valid results and must not cause any harm that cannot be justified. The Act stipulates, for example, that high risk systems have to be tested 'against preliminary defined metrics and probabilistic thresholds that are appropriate to the intended purpose',[165] that training, validation, and testing data must be 'relevant, representative, free of errors and complete' and shall have the 'appropriate statistical properties',[166] that data governance must include bias monitoring,[167] that the systems achieve 'in the light of their intended purpose, an appropriate level of accuracy'[168] and that 'levels of accuracy and the relevant accuracy metrics' have to be declared in the instructions of use.[169] As many of the AI systems known for their discrimination risks are classified as 'high risk'[170] or may be classified accordingly by the Commission in the future,[171] this is already a good start.

[164] EU Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 21st April 2021, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206; *Cf.* T Burri, Chapter 7, in this volume.

[165] EU Artificial Intelligence Act, Article 9(7).

[166] EU Artificial Intelligence Act, Article 10(3).

[167] EU Artificial Intelligence Act, Article 10(2)(f) and (5).

[168] EU Artificial Intelligence Act, Article 15(1).

[169] EU Artificial Intelligence Act, Article 15(2).

[170] For example those used for predicting job performance, creditworthiness, or crime. See EU Artificial Intelligence Act, Annex III.

[171] EU Artificial Intelligence Act, Article 7.

PART V

Responsible Data Governance

16

Artificial Intelligence and the Right to Data Protection

Ralf Poscher

In respect of technological advancement, the law often comes into play merely as an external restriction. That is, lawyers are asked whether a given technology is consistent with existing legal regulations or to evaluate its foreseeable liability risks. As a legal researcher, my interest is the exact opposite: how do new technologies influence our legal framework, concepts, and doctrinal constructions? This contribution shows how Artificial Intelligence (AI) challenges the traditional understanding of the right to data protection and presents an outline of an alternative conception, one that better deals with emerging AI technologies.

I. TRADITIONAL CONCEPT OF THE RIGHT TO DATA PROTECTION

In the early stages of its data protection jurisprudence, the German Federal Constitutional Court took a leading role in establishing the right to data protection, not only in Germany, but also in the European context.[1] In the beginning, it linked the 'right to informational self-determination' to a kind of property rights conception of personal data.[2] The Court explained that every individual has a 'right to determine himself, when and in which boundaries personal data is disseminated'[3] – just as an owner has the right to determine herself when she allows someone to use her property.[4] This idea, which is already illusory in the analog world, has often been ridiculed as naive in our contemporary, technologically interconnected and socially networked reality, in which a vast spectrum of personal data is disseminated and exchanged at all levels almost all of the time.[5] Data simply does not possess the kind of exclusivity to justify parallels

[1] M Albers, 'Realizing the Complexity of Data Protection' in S Gutwirth, R Leenes, and P De Hert (eds), *Reloading Data Protection* (2014) 217 (hereafter Albers, 'Complexity'); K Vogelsang, *Grundrecht auf Informationelle Selbstbestimmung?* (1987) 39–88.

[2] There is a certain parallel between this conceptualization of the right to privacy and its scope under the US Supreme Court's early Fourth Amendment jurisprudence: the Supreme Court, until *Katz v United States* 389 US 347 [1967], applied the Fourth Amendment only to the search and seizure of a citizen's personal property and effects (see, e.g., *Olmstead v United States* 277 US 438 [1928]) and was thus tied in substance to a property right.

[3] BVerfGE 65, 1 (42) (BVerfG *1 BvR 209/83*): ‚Befugnis des Einzelnen, grundsätzlich selbst zu entscheiden, wann und innerhalb welcher Grenzen persönliche Lebenssachverhalte offenbart werden.'

[4] Albers, 'Complexity' (n 1) 219.

[5] M Albers, 'Information als neue Dimension im Recht' (2002) 33 *Rechtstheorie* 61 (81) (hereafter Albers, 'Information'); K Ladeur, 'Das Recht auf Informationelle Selbstbestimmung: Eine Juristische Fehlkonstruktion?'(2009) 62 *DÖV* 45 (46–47).

with property ownership.[6] The German Constitutional Court seems to have recognized this. And while the Court has not explicitly revoked the property-like formula, it has made decreasing use of it, and in more recent decisions, has not referred to it at all.[7]

Even if everyone can agree that the right to data protection is, in substance, not akin to a property interest in one's personal data, the right to data protection is formally handled as if it were a property right. In the same way that any non-consensual use of one's property by someone else is regarded a property rights infringement, any non-consensual use – gathering, storage, processing, and transmission – of personal data is viewed as an infringement of the right to data protection. This formal conception of data protection is not only still prevalent in the German context, but the European Court of Justice (ECJ) perceives the right to data protection under Article 8 of the Charter of Fundamental Rights of the European Union (CFR) in much the same way. In one of its latest decisions, the ECJ confirmed that data retention as such constitutes an infringement irrespective of substantive inconveniences for the persons concerned:

> It should be made clear, in that regard, that the retention of traffic and location data constitutes, in itself, … an interference with the fundamental rights to respect for private life and the protection of personal data, enshrined in Articles 7 and 8 of the Charter, irrespective of whether the information in question relating to private life is sensitive or whether the persons concerned have been inconvenienced in any way on account of that interference.[8]

According to the traditional perspective, each and every processing of personal data infringes the respective right – just as the use of physical property would be an infringement of the property right.[9] For instance, if my name, license plate, or phone number is registered, this counts as an infringement; if they are stored in a database, this counts as another infringement; and if they are combined with other personal data, such as location data, this counts as yet another infringement.[10] Even though the right to data protection is not regarded as a property right, its formal structure still corresponds with that of a property right.

This conceptual approach is a mixed blessing. On the one hand, it provides a very analytic approach to the data processing in question. On the other hand, the idea of millions of fundamental rights infringements occurring in split seconds by CPUs processing personal data seems a rather exaggerated way of conceptualizing the actual problems at hand. Nevertheless, modern forms of data collection are still conceptualized in this way, including automated license plate recognition, whereby an initial infringement occurs by using scanners to collect license plate information and another infringement by checking this information against stolen car databases,[11] etc.

[6] Cf. J Fairfield and C Engel, 'Privacy as a Public Good' in RA Miller (ed), *Privacy and Power: A Transatlantic Dialogue in the Shadow of the NSA-Affair* (2017).

[7] E.g., BVerfGE 120, 351 (360) (BVerfG *1 BvR 2388/03*); BVerfGE 120, 378 (397–398) (BVerfG *1 BvR 2074/05*).

[8] CJEU, Joined Cases C-511/18, C-512/18, and C-520/18 *La Quadrature du Net and Others v Premier ministre and Others* (6 October 2020), para 115 (hereafter CJEU, *La Quadrature du Net*).

[9] Albers, 'Complexity' (n 1) 219.

[10] BVerfGE 100, 313 (366) (BVerfG *1 BvR 2226/94*); BVerfGE 115, 320 (343–344) (BVerfG *1 BvR 518/02*); BVerfGE 125, 260 (310) (BVerfG *1 BvR 256, 263, 586/08*); BVerfGE 130, 151 (184) (BVerfG *1 BvR 1299/05*); BVerfGE 150, 244 (265–266) (BVerfG *1 BvR 142/15*).

[11] BVerfGE 120, 378 (400–401) (BVerfG *1 BvR 1254/05*); BVerfGE 150, 244 (266) (BVerfG *1 BvR 142/15*).

II. THE INTRANSPARENCY CHALLENGE OF AI

AI technology is driven by self-learning mechanisms.[12] These self-learning mechanisms can adapt their programmed algorithms reacting to the data input.[13] Importantly, while the algorithms may be transparent to their designers,[14] after the system has cycled through hundreds, thousands, or even millions of recursive, self-programming patterns, even the system programmers will no longer know which type of data was processed in which way, which inferences were drawn from which data correlations, and how certain data have been weighted.[15]

The self-adaptive 'behavior' of at least certain types of AI technologies leads to a lack of transparency. This phenomenon is often referred to as the black box issue of AI technologies.[16] Why is this a problem for the traditional approach to evaluating data protection?

The analytical approach is based on the justification of each and every processing of personal data. In AI systems, however, we do not know which individual personal data have been used and how many times they have been processed and cross-analyzed with what types of other data.[17] It is thus impossible to apply the analytical approach to determine whether, how many, and what kind of infringements on a thus conceived right to data protection occurred. AI's lack of transparency seems to rule this out. Thus, AI creates problems for the traditional understanding and treatment of the right to data protection due to its lack of transparency.[18] These issues are mirrored in the transparency requirements of the General Data Protection Regulation, which rests very much on the traditional conception of the fundamental right to data protection.[19]

III. THE ALTERNATIVE MODEL: A NO-RIGHT THESIS

The alternative conceptualization of the right to data protection that I would like to suggest consists of two parts.[20] The first part sounds radical, revisionary, and destructive; the second part resolves the tension created by a proposal that is doctrinally mundane but shifts the perspective

[12] H Surden, 'Machine Learning and Law' (2014) 89 *Washington L Rev* 87 (88–90) (hereafter Surden, 'Machine Learning'); W Hoffmann-Riem, 'Verhaltenssteuerung durch Algorithmen – Eine Herausforderung für das Recht' (2017) 142 *AöR* 3 (hereafter Hoffmann-Riem, 'Verhaltenssteuerung'); W Hoffmann-Riem, 'Artificial Intelligence as a Challenge for Law and Regulation' in T Wischmeyer and T Rademacher (eds), *Regulating Artificial Intelligence* (2020) 3 (hereafter Hoffmann-Riem, 'Artificial Intelligence').

[13] Surden, 'Machine Learning' (n 12) 93.

[14] Hoffmann-Riem, 'Verhaltenssteuerung' (n 12) 30.

[15] Hoffmann-Riem, 'Artificial Intelligence' (n 12), 17; Hoffmann-Riem, 'Verhaltenssteuerung' (n 12) 29; N Marsch, 'Artificial Intelligence and the Fundamental Right to Data Protection' in T Wischmeyer and T Rademacher (eds), *Regulating Artificial Intelligence* (2020) 36 (hereafter Marsch, 'Artificial Intelligence'); T Wischmeyer, 'Artificial Intelligence and Transparency: Opening the Black Box' in T Wischmeyer and T Rademacher (eds), *Regulating Artificial Intelligence* (2020) 81 (hereafter Wischmeyer, 'Artificial Intelligence').

[16] Hoffmann-Riem, 'Verhaltenssteuerung' (n 12) 29; Marsch, 'Artificial Intelligence' (n 15) 36; Wischmeyer, 'Artificial Intelligence' (n 15) 80.

[17] Cf. Albers, 'Complexity' (n 1) 221: 'The entire approach is guided by the idea that courses of action and decision-making processes could be almost completely foreseen, planned and steered by legal means'; Marsch, 'Artificial Intelligence' (n 15) 39.

[18] Marsch, 'Artificial Intelligence' (n 15) 36.

[19] On the specifics of the transparency requirements generally stated in Articles 5(1)(a) alt. 3 GDPR and the issues the cause for the use of AI-technologies, B Paal, Chapter 17 in this volume.

[20] For a more general discussion of this alternative account, see R Poscher, 'Die Zukunft der Informationellen Selbstbestimmung als Recht auf Abwehr von Grundrechtsgefährdungen' in H Gander and others (eds), *Resilienz in der offenen Gesellschaft* (2012) 171–179; R Poscher, 'The Right to Data Protection' in RA Miller (ed), *Privacy and Power: A Transatlantic Dialogue in the Shadow of the NSA-Affair* (2017) 129–141.

on data protection rights substantially. Among other advantages, the proposed shift in perspective could render the right to data protection more suitable for handling issues arising from AI.

The first part is a no-right-thesis. It contends that there is no fundamental right to data protection. That is, the right to data protection is not a right of its own standing. This explains why the ongoing quest for a viable candidate as the proper object of the right to data protection has been futile.[21] Article 8 CFR, which seems to guarantee the right to data protection as an independent fundamental right, rests on the misunderstanding that the fundamental rights developments in various jurisdictions, namely also in the jurisdiction of the German Federal Constitutional Court, have created a new, substantive fundamental right with personal data as its object. There is no such new, substantive fundamental right. This, however, does not mean that there is no fundamental rights protection against the collection, storage, processing, and dissemination of personal data. Yet data protection does not take the form of a new fundamental right – property-like or otherwise.

The second part of the thesis reconstructs the 'right' by shifting the focus to already existing fundamental rights. Data protection is provided by all of the existing fundamental rights, which can all be affected by the collection, storage, processing, and dissemination of personal data.[22] In his instructive article 'A Taxonomy of Privacy', *Daniel Solove* developed a whole taxonomy of possible harms that can be caused by data collection.[23] They include the loss of life and liberty, infringements on property interests and the freedom of expression, violations of privacy, and denials of due process guarantees. It is easy to see how the dissemination of personal finance information can lead to the loss of property. He cites tragic cases, which have even led to a loss of life, such as when a stalker was handed the address of his victim by public authorities – data he used to locate and kill her.[24] *Solove's* list suggests that the essence of data protection cannot be pinned down to merely a single liberty or equality interest but instead potentially involves every fundamental right. Understood correctly, the right to data protection consists in the protection that all fundamental rights afford to all the liberty and equality interests that might be affected by the collection, storage, processing, and dissemination of personal data.

The way in which fundamental rights protect against the misuse of personal data relies on doctrinally expanding the concept of rights infringement. Fundamental rights usually protect against actual infringements. For example, the state encroaches upon your right of personal freedom if you are incarcerated, your right to freedom of assembly is infringed when your meeting is prohibited or dispersed by the police, and your freedom of expression is violated when you are prevented from expressing your political views. Usually, however, fundamental rights do not protect against the purely abstract danger that the police might incarcerate you, might disperse your assembly, or might censor your views. You cannot go to the courts claiming that certain police behavioral patterns increase the danger that they might violate your right to assembly. The courts would generally say that you have to wait until they either already do so or are in the concrete process of doing so. In some cases, your fundamental rights might already protect you if there is a concrete danger that such infringements are about to take place, so that

[21] C Gusy, 'Informationelle Selbstbestimmung und Datenschutz: Fortführung oder Neuanfang?' (2000) 83 *KritV* 52, 56–63; K Ladeur, 'Das Recht auf Informationelle Selbstbestimmung: Eine Juristische Fehlkonstruktion?' (2009) 62 *DÖV* 45, 47–50.

[22] N Marsch, *Das Europäische Datenschutzgrundrecht* (2018), 92 (hereafter Marsch, 'Datenschutzgrundrecht').

[23] DJ Solove, 'A Taxonomy of Privacy' (2006) 154 *U Pennsylvania L Rev* 477; see also DJ Solove, '"I've Got Nothing to Hide" and Other Misunderstandings of Privacy' (2007) 44 *San Diego L Rev* 745, 764–772 (hereafter Solove, 'Misunderstandings of Privacy').

[24] Solove, 'Misunderstandings of Privacy' (n 23) 768.

you do not have to suffer the infringement in the first place if it were to violate your rights.[25] These cases, however, are exceptions.

The right to data protection works differently. What is unique about data protection is its generally preemptive character. It already protects against the abstract dangers involved in the collection, storage, and processing of personal data.[26] Data protection preemptively protects against violations of liberty or equality interests that are potentially connected to *using* personal data.[27] The collection, aggregation, and processing of data as such does no harm.[28] This has often been expressed in conjunction with the idea that data needs to become information in certain contexts before it gains relevance.[29] It is only the use of data in certain contexts that might involve a violation of liberty or equality interests. The collection of personal data on political or religious convictions of citizens by the state is generally prohibited, for example, because of the potential that it could be misused to discriminate against political or religious groups. Data protection demands a justification for the collection of personal data, even if such misuse is only an abstract danger.[30] It does not require concrete evidence that such misuse took place, or even that such misuse is about to take place. The right to data protection systematically enhances every other fundamental right already in place to protect against the abstract dangers that accompany collecting and processing personal data.[31]

A closer look at the court practice regarding the right to data protection reveals that, despite appearances, courts neither treat the right to data protection as a right on its own but instead associate it with different fundamental rights, depending on the context and the interest affected.[32] Even at the birth of the right to data protection in Germany, in the famous "Volkszählungs-Urteil" (census decision), the examples the court gave to underline the necessity for a new fundamental right to 'informational self-determination' included a panoply of

[25] See BVerfGE 51, 324 (BVerfG 2 *BvR 1060/78*), in which the Court saw it as an infringement of the right to physical integrity to proceed with a criminal trial if the defendant runs the risk of suffering a heart attack during the trial; cf. also BVerfGE 17, 108 (BVerfG *1 BvR 542/62*) (high-risk medical procedure – lumbar puncture – with the aim of determining criminal accountability for a misdemeanor); BVerfGE 52, 214 (220) (BVerfG *1 BvR 614/79*)(eviction of a suicidal tenant) and R Poscher, *Grundrechte als Abwehrrechte* (2003) 388–390 (hereafter Poscher, 'Abwehrrechte').

[26] Cf. Marsch, 'Datenschutzgrundrecht' (n 22) 109, with a focus on the internal peace of mind of deciding on one's exercise of fundamental rights.

[27] E.g., the collection of comprehensive data in the course of a nationwide census is not in itself an imminent threat, but it is dangerous because of the potential (mis-)use of the masses of the gathered mass data, cf. BVerfG 65, 1 (BVerfG *1 BvR 209/8*); the collection of data for an anti-terrorism or anti-Nazi database is problematic because of *potential* negative impacts for those mentioned in it, cf. BVerfGE 133, 277 (331–332) (BVerfG *1 BvR 1215/07*).

[28] Albers, 'Complexity' (n 1) 225.

[29] M Albers, 'Zur Neukonzeption des Grundrechtlichen „Daten"Schutzes' in A Haratsch and others (eds), *Herausforderungen an das Recht der Informationsgesellschaft* (1996) 121–23, 131–33; Albers, 'Information' (n 5) 75; M Albers, *Informationelle Selbstbestimmung* (2005) 87–148; M Albers, 'Umgang mit Personenbezogenen Informationen und Daten' in W Hoffmann-Riem, E Schmidt-Aßmann and A Voßkuhle (eds) *Grundlagen des Verwaltungsrechts* (2nd ed. 2012) 7–28; G Britz, 'Informationelle Selbstbestimmung Zwischen Rechtswissenschaftlicher Grundsatzkritik und Beharren des Bundesverfassungsgerichts' in W Hoffmann-Riem (ed), *Offene Rechtswissenschaft* (2010) 566–568 (hereafter Britz, 'Informationelle Selbstbestimmung'); Albers, 'Complexity' (n 1) 222–224.

[30] Cf. the examples mentioned in note 27. This pre-emptive protection against state action is not to be confused with the duties to protect against *unlawful* infringements of liberty interests by *third parties*, cf. Poscher, 'Abwehrrechte' (n 25) 380–387 on the duty to protect under the German Basic Law. As far as such duties to protect are accepted, data protection would also address pre-emptive dimensions of these duties.

[31] Cf. J Masing, 'Datenschutz – ein unterentwickeltes oder überzogenes Grundrecht?' (2014) *RDV* 3 (4); Marsch, 'Datenschutzgrundrecht' (n 22) 109–110; T Rademacher, 'Predictive Policing im Deutschen Polizeirecht' (2017) 142 *AöR* 366 (402); Marsch, 'Artificial Intelligence' (n 15) 40.

[32] Cf. Britz, 'Informationelle Selbstbestimmung' (n 29) 571, 573, who first characterized the German right to informational self-determination as an 'accessory' right.

fundamental rights, such as the right to assembly.[33] In an unusual process of constitutional migration, the court pointed to the 'chilling effects' the collection of data on assembly participation could have for bearers of that right,[34] as they were first discussed by the US Supreme Court.[35] The German Federal Court drew on an idea developed by the US Supreme Court to create a data protection right that was never accepted by the latter. Be that as it may, even in its constitutional birth certificate, data protection is not put forth as a right on its own but associated with various substantive fundamental rights, such as the right to assembly.

Further evidence of the idea that personal data is not the object of a substantive stand-alone right is provided by the fact that data protection does not seem to stand by itself, even in a jurisdiction in which it is explicitly guaranteed. Article 8 CFR explicitly guarantees a right to data protection. In the jurisprudence of the Court of Justice of the European Union, however, it is always cited in conjunction with another right.[36] The right to data protection needs another right in order to provide for a substantive interest – usually the right to privacy,[37] but sometimes also other rights, such as free speech.[38] Thus, even when data protection is codified as an explicit, independent fundamental right, as it is in the Charter, it is nevertheless regarded as an accessory to other more substantive fundamental rights.[39] This is odd if the right to data protection is taken at face value as a substantive right on its own but only natural if taken as a general enhancement of other fundamental rights.

IV. THE IMPLICATION FOR THE LEGAL PERSPECTIVE ON AI

If the right to data protection consists in a general enhancement of, potentially, every fundamental right in order to already confront the abstract dangers to the liberty and equality interests they protect, it becomes clear how personal data processing systems must be evaluated. They have to be evaluated against the background of the question: to what extent does a certain form of data collection and processing system pose an abstract danger for the exercise of what type of fundamental right? Looking at data collection issues in this way has important implications – including for the legal evaluation of AI technologies.

1. *Refocusing on Substantive Liberty and Equality Interests*

First, the alternative conception allows us to rid ourselves of a formalistic and hollow understanding of data protection. It helps us to refocus on the substantive issues at stake. For many people, the purely formal idea that some type of right is always infringed when a piece of personal information has been processed, meaning that they have to sign a consent agreement or

[33] BVerfGE 65, 1 (43) (BVerfG *1 BvR 209/83*).
[34] BVerfGE 65, 1 (43) (BVerfG *1 BvR 209/83*).
[35] *Wieman v Updegraff* 344 US 183 (1952), para 195.
[36] CJEU, Joined Cases C-92/09 and C-93/09 *Schecke and Eifert v Hesse* [2010] ECR I-11063, para 47; CJEU, Joined Cases C-293/12 and C-594/12 *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources, Minister for Justice, Equality and Law Reform, The Commissioner of the Garda Siochana, Ireland and the Attorney General and Kärntner Landesregierung, Michael Seitlinger, Christof Tschohl and Others* (8 April 2014), para 53 (hereafter CJEU, *Digital Rights Ireland*); CJEU, Case C-362/14 *Maximillian Schrems v Data Protection Commissioner* (6 October 2015), para 78; CJEU, Case C-311/18 *Data Protection Commissioner v Facebook Ireland Limited and Maximilian Schrems* (16 July 2020), para 168.
[37] CJEU, *Digital Rights Ireland* (n 36) para 37; CJEU, *La Quadrature du Net* (n 8) para 115.
[38] CJEU, *Digital Rights Ireland* (n 36) para 28; CJEU, *La Quadrature du Net* (n 8) para 118; CJEU, Case C-623/17 *Privacy International v Secretary of State for Foreign and Commonwealth Affairs and Others* (6 October 2020), para 72.
[39] Marsch, 'Datenschutzgrundrecht' (n 22) 132–133.

click a button, has become formalistic and stale in the context of data protection regulation. The connection to the actual issues that are connected with data processing has been lost. For example, during my time as vice dean of our law faculty, I attempted to obtain the addresses of our faculty alumni from the university's alumni network. The request was denied because it would constitute an infringement of the data protection right of the alumni. The alumni network did not have the written consent of its members to justify this infringement. As absurd as this might seem, this line of argument is the only correct one for the traditional, formal approach to data protection. Addresses are personal data and any transfer of this personal data is an infringement of the formal right to data protection, which has to be justified either by consent or by a specific statute – both of which were lacking. This is, however, a purely formal perspective. Our alumni would probably be surprised to know that the faculty at which they studied for years, which handed them their law degrees, and which paved the road to their legal career does not know that it is their alma mater. There is no risk involved for any of their fundamental rights when the faculty receives their address information from the alumni network of the very same university. An approach that discards the idea that there is a formal right to data protection, but asks which substantive fundamental rights positions are at stake, can resubstantialize the right to data protection. This also holds for AI systems: the question would not be what type of data is processed when and how but instead what kind of substantive, fundamental right position is endangered by the AI system.

2. *The Threshold of Everyday Digital Life Risks*

Second, refocusing on the abstract danger for concrete, substantive interests protected by fundamental rights allows for a discussion on thresholds. Also, in the analog world, the law does not react to each and every risk that is associated with modern society. Not every abstract risk exceeds the threshold of a fundamental rights infringement. There are general life risks that are legally moot. In extreme weather, even healthy trees in the city park carry the abstract risk that they might topple, fall, and cause considerable damage to property or even to life and limb. Courts, however, have consistently held that this abstract danger does not allow for public security measures or civil claims to chop down healthy trees.[40] They consider it part of everyday life risks that we all have to live with if we stroll in public parks or use public paths.

The threshold for everyday life risks holds in the analog world and should hold in the digital world, too. In our digital society, we have to come to grips with a – probably dynamic – threshold of everyday digital life risks that do not constitute a fundamental rights infringement, even though personal data have been stored or processed. On one of my last visits to my physician, I was asked to sign a form that would allow his assistants to use my name, which is stored in their digital patient records, in order to call me from the waiting room when the doctor is ready to see me. The form cited the proper articles of the, at the time, newly released General Data Protection Regulation of the European Union (Articles 6(1)(a) and 9(2)(a)). There might be occasions where there is some risk involved in letting other patients know my name. If the physician in question were an oncologist, it might lead to people spreading the rumor that I have a terminal illness. This might find its way to my employer at a time when my contract is up for an extension. So, there can indeed be some risk involved. We have, however, always accepted this risk – also in a purely analog world – as one that comes with the visit of physicians, just as we have accepted the risk of healthy trees being uprooted by a storm and damaging our houses, cars,

[40] VG Minden (11 K 1662/05) [2005], para 32.

or even hurting ourselves. As we have accepted everyday life risks in the analog world, we have to accept everyday digital life risks in the digital world.

For AI technologies, this could mean that they can be designed and implemented in a way that they remain below the everyday digital life risk threshold. When an AI system uses anonymized personal data, there is always a risk that the data will be deanonymized. If sufficient safeguards against deanonymization are installed in the system, however, they may lower the risk to such a degree that it does not surpass the level of our everyday digital life risk. This may be the case if the AI system uses data aggregation for planning purposes or resource management, which do not threaten substantive individual rights positions. An example of a non-AI application is the German Corona-Warn-App, which is designed in such a way as to avoid centralized storage of personal data and thus poses almost no risk of abuse.

3. *A Systemic Perspective*

Third, the alternative approach implies a more systemic perspective on data collection and data processing measures. It allows us to step back from the idea that each and every instance of personal data processing constitutes an infringement of a fundamental right. If data protection is understood as protection against abstract dangers, then we do not have to look at the individual instances of data processing. Instead, we can concentrate on the data processing system and its context in order to evaluate the abstract danger it poses.

Unlike the traditional approach, focusing on abstract dangers for substantive fundamental rights that are connected with AI technologies does not require the full transparency of the AI system. The alternative approach does not require exact knowledge of when and how what kind of data is processed. What it needs, however, is a risk analysis and an evaluation of the risk reduction, management, correction, and compensation measures attuned to the specific context of use.[41] It requires regulation on how false positives and negatives are managed in the interaction between AI and human decision makers. At the time of our conference, the *New York Times* reported on the first AI-based arrest generated by a false positive of facial recognition software.[42] As discussed in the report, to rely solely on AI-based facial recognition software for arrests seems unacceptable given the failure rate of such systems. Legal regulation has to counterbalance the risks stemming from AI by forcing the police to corroborate AI results with additional evidence. A fundamental rights analysis of the facial recognition software should include an evaluation not only of the technology alone but also of the entire sociotechnological arrangement in the light of *habeas corpus* rights and the abstract dangers for the right to personal liberty that come with it. The actual cases, however, are not about some formal right to data protection but about substantive rights, such as the right to liberty or the right against racial discrimination, and the dangers AI technologies pose for these rights.

For AI technologies, the differences between the traditional approach and the suggested approach regarding the right to data protection are similar to differences in the scientific approach to, and the description of, the systems as such. Whereas traditionally the approach to, and the description of, computational systems has been very much dominated by computer sciences, there is a developing trend to approach AI systems – especially because of their lack of informational transparency – with a more holistic intradisciplinary methodology. AI systems are

[41] Cf. Albers, 'Complexity' (n 1) 232, who draws a parallel to risk management in environmental law.

[42] K Hill, 'Wrongfully Accused by an Algorithm' *New York Times* (24 June 2020). nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

studied in their deployment context with behavioral methodologies which are not so much focused on the inner informational workings of the systems but on their output and their effects in a concrete environment.[43] The traditional approach tends toward a more technical, informational analysis of AI systems, which is significantly hampered by the black box phenomenon. The shift to the substantive rights perspective would lean toward a more behavioral approach to AI. The law would not have to delve into the computational intricacies of when and how what type of personal data is processed. It could take a step back and access how an AI system 'behaves' in the concrete sociotechnological setting it is employed in and what type of risks it generates for which substantive fundamental rights.

V. CONCLUSION

From a doctrinal, fundamental rights perspective, AI could have a negative and a positive implication. The negative implication pertains to the traditional conceptualization of data protection as an independent fundamental right on its own. The traditional formal model, which focuses on each and every processing of personal data as a fundamental rights infringement could be on a collision course with AI's technological development. AI systems do not provide the kind of transparency that would be necessary to stay true to the traditional approach. The positive implication pertains to the alternative model I have been suggesting for some time. The difficulties AI may pose for the traditional conceptualization of the right to data protection could generate some wind beneath the wings of the alternative conception, which seems better equipped to handle AI's black box challenge with its more systemic and behavioral approach. The alternative model might seem quite revisionary, but it holds the promise of redirecting data protection toward the substantive fundamental rights issues at stake – also, but not only, with respect to AI technologies.

[43] An overview on this emerging field in I Rahwan and others, 'Machine behaviour' (2019) 568 *Nature* 477 (481–482).

17

Artificial Intelligence as a Challenge for Data Protection Law

And Vice Versa

*Boris P. Paal**

I. INTRODUCTION

Artificial Intelligence (AI) as an area of research within the field of computer science concerns itself with the functioning of autonomous systems and, as such, not only affects almost all areas of modern life in the age of digitisation but has also – and for good reasons – become a focal point within both academic and political discourse.[1] AI scenarios are mainly driven and determined by the availability and evaluation of data. In other words, AI goes hand in hand with what may be referred to as an enormous 'appetite for data'. Thus, the accumulation of relevant (personal or non-personal) data regularly constitutes a key factor for AI-related issues. The collected personal data may then be used to create (personality) profiles as well as to make predictions and recommendations with regard to individualised services and offers. In addition, non-personal data may be used for the analysis and maintenance of products. The applications and business models based on the collection of data are employed in both the private and public sector. The current and potential fields of application for AI are as diverse and numerous as the reactions thereto, ranging from optimism to serious concerns – oftentimes referring to a potential 'reign of the machines'. However, there is a general consensus regarding the fact that the development and use of AI technologies will have significant impact on the state, society, and economy. For instance, the use of such applications may greatly influence the protection of personal rights and privacy, because the development of AI technologies regularly requires the collection of personal data and the processing thereof. This chapter will focus on and examine provisions concerning the handling of personal data as set out in the European Union's General Data Protection Regulation (GDPR)[2] which entered into force on 24 May 2016 and has been applicable since 25 May 2018.

The prerequisites and applications of AI on one hand and the regulatory requirements stipulated by the GDPR on the other, give rise to a number of complicated, multi-sided tensions and conflicts. While the development of AI is highly dependent on the access to large amounts

* Transcript of a presentation held at the Conference Global Perspectives on Responsible AI 2020 in Freiburg on June 26, 2020. The presentation form was maintained for the most parts. Fundamental considerations of this paper are also published in B Paal, 'Spannungsverhältnis von KI und Datenschutzrecht' in M Kaulartz and T Braegelmann (eds), *Rechtshandbuch Artificial Intelligence und Machine Learning* (2020) 427–444.

[1] On defining AI see for example J Kaplan, *Artificial Intelligence* (2016) 1 *et seq*.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

of data (i.e. big data), this access is subject to substantial limitations imposed by the data protection law regime. These restrictions mainly apply to scenarios concerning personal (instead of non-personal) data and primarily stem from the GDPR's preventive prohibition subject to authorisation[3] and its general principles relating to the processing of personal data.[4] One of the most fundamental problems which arises in connection with big data is referred to as 'small privacy'. This term alludes to the inherent conflict between two objectives pursued by data protection law, the comprehensive protection of privacy and personal rights on the one hand and the facilitation of an effective and competitive data economy on the other. The tension arising from this conflict is further illustrated by Article 1 GDPR, according to which the Regulation contains provisions to promote both the protection of natural persons with regard to the processing of personal data and the free movement of such data. An instrument intended to facilitate an appropriate balance between the protection of personal data and seemingly contradictory economic interests may be seen in the users' data sovereignty.[5]

At this point, it should be noted that the GDPR does not (or, if at all, only marginally) address the implications of AI for data protection law. Thus, in order to be applied to individual cases and to specific issues arising in connection with AI, the general provisions of the GDPR need to be construed. This may oftentimes lead to substantial legal uncertainties, especially when considering the vague wording, unclear exemptions, and considerable administrative discretion provided by the GDPR. The aforementioned uncertainties may not only impede innovation but may also give rise to a number of issues concerning the (legal) accountability for AI, for instance, in connection with the so-called black-box-phenomenon[6] regularly encountered when dealing with self-learning AI systems (i.e. deep or machine learning).

II. AI AND PRINCIPLES RELATING TO THE PROCESSING OF DATA

The development and use of AI may potentially conflict with almost all principles concerning the processing of data as enshrined in the GDPR. In fact, the paradigms of data processing in an AI-context are very difficult, if not impossible, to reconcile with the traditional principles of data protection. The complex and multi-layered legal issues resulting from this contradiction are first and foremost attributable to the fact that AI scenarios were not (sufficiently) taken into account during the drafting of the GDPR. This raises the question of whether and to what extent AI scenarios can be adequately addressed and dealt with under the existing legal regime by utilising the available technical framework and by interpreting the relevant provisions accordingly. Where the utilisation of such measures and, consequently, the application of the law and the compliance[7] with the principles of data protection is not possible, it has to be assessed whether there are any other options to adapt or to amend the existing legal framework.[8]

The aforementioned data protection issues have their roots in the general principles of data protection. Hence, in order to fully comprehend the (binding) provisions that a 'controller' in the sense of the GDPR must observe when processing data, it is necessary to take a closer look at these principles. This is especially important considering the very prominent role of the legal

[3] Cf. GDPR, Article 6(1).
[4] Cf. GDPR, Article 5.
[5] On data sovereignty see for example PL Krüger, 'Datensouveränität und Digitalisierung' (2016) *ZRP* 190.
[6] On the 'black box-phenomenon' see for example F Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information* (2015).
[7] For this see the following Section III.
[8] For this see the following Section IV.

framework in nearly all AI scenarios. The addressee of the principles relating to the processing of personal data laid down in Article 5(1) GDPR, is responsible for the adherence thereto and must, as required by the principle of accountability, be able to provide evidence for its compliance therewith.[9] The obligations set out in Article 5(1) GDPR range from the lawfulness, fairness, and transparency of data processing as well as the adherence to and compatibility with privileged purposes (purpose limitation) to the principle of data minimisation, accuracy, storage limitation, as well as integrity and confidentiality.[10] Beyond the scope of the present analysis in this chapter lie questions concerning conflicts of law and the lawfulness of data transfer in non-EU Member States, although these constellations are likely to become increasingly important in legal practice especially in light of the growing importance of so-called cloud-solutions[11].

1. *Transparency*

In accordance with Article 5(1)(a) alt. 3 GDPR, personal data must be 'processed [...] in a transparent manner in relation to the data subject'. These transparency requirements are of particular importance for matters relating to AI. As set out in Recital 39 of the GDPR, the principle of transparency

> requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used. That principle concerns, in particular, information to the data subjects on the identity of the controller and the purposes of the processing and further information to ensure fair and transparent processing in respect of the natural persons concerned and their right to obtain confirmation and communication of personal data concerning them which are being processed.[12]

These transparency requirements are specified by the provisions contained in Articles 12-15 GDPR, which stipulate the controllers' obligation to provide information and to grant access to personal data. They are further accompanied by the obligation to implement appropriate technical and organisational measures.[13] Moreover, Article 12(1) sentence 1 GDPR requires the controller to provide the data subject with any information and communication 'in a concise, transparent, intelligible and easily accessible form, using clear and plain language'. Especially with regard to issues relating to AI, the implementation of these requirements is very likely to present responsible controllers with a very complex and onerous task.

In an AI scenario, it will often be difficult to state and substantiate the specific purposes for any given data analysis in advance. Controllers may also face enormous difficulty when tasked with presenting the effects that such an analysis could have on the individual data subject in a sufficiently transparent manner. In fact, the very nature of self-learning AI which operates with unknown (or even inexplicable) variables seems to oppose any attempt to present and provide any transparent information.[14] In addition, the aforementioned 'black-box-phenomena' may

[9] Cf. GDPR, Article 5(2).

[10] GDPR, Article 5(1)(a)–(f).

[11] On GDPR and the cloud see J Krystelik, 'With GDPR, Preparation Is Everything' (2017) *Computer Fraud & Security* 5 (7).

[12] See also Article 29 Data Protection Working Party, 'Guidelines on Transparency under Regulation 2016/679' WP 260 rev.01.

[13] Cf. GDPR, Recital 78. For this see the following Section III 3.

[14] L Mitrou, 'Data Protection, Artificial Intelligence and Cognitive Services: Is the GDPR "Artificial Intelligence-Proof"?' (2018) Tech Report commissioned by Microsoft, 58 https://ssrn.com/abstract=3386914 (hereafter Mitrou, 'Data Protection').

occur if, for instance, artificial neural networks on so-called hidden layers[15] restrict or even prohibit the traceability of the respective software-processes. Thus, on the one hand, it may be difficult to break down the complex and complicated AI analyses and data collection processes into 'concise, transparent, intelligible and easily accessible' terms that the affected data subject can understand. On the other hand, the lack of transparency is an inherent feature and characteristic of self-learning, autonomous AI technologies.[16] Furthermore, these restrictions on transparency also come into play when considering potential justifications for the processing of data. This is particularly relevant where the justification is based on the data subject's consent as this (also) requires an informed indication of the subject's agreement.[17]

However, according to the principles of the GDPR, even controllers who use systems of AI and, thus, carry out extensive analyses of huge amounts of data of different origins, should have the (realistic) possibility to process data in a manner which allows them to adequately inform the subject about the nature and origin of the processed data. Further difficulties are likely to arise in situations where personal data are generated in the course of analyses or as a result of combinations of originally non-personal data. Because, in this case, the legally relevant collection of data is to be found in the analysis, it is difficult if not impossible to pinpoint the data's initial origin and source. In such constellations, it should, thus, be assumed that the responsible controller is permitted to merely provide general information, for instance by naming the source of the data stock or the systems utilised to process the data in addition to the means used for their collection. In this context, it also has to be emphasised that the obligation to inform the data subject as set out in Article 14(5)(b) GDPR may be waived if the provision of such information would be disproportionally onerous. The applicability of this waiver must be determined by balancing the controller's efforts required for the provision of information with the data subject's right and interest to be informed. The outcome of this (case-by-case) balancing process in big-data-situations – not only in the context of AI – will largely depend on the effects that the data analysis and processing have on the subject's fundamental rights, as well as on the nature and degree of risks that arise in connection thereto. For the purposes of such an assessment, the principle of transparency should extend beyond the actual data processing procedures to include the underlying technical systematics and the decision-making systems employed by the (responsible) controller.

2. *Automated Decisions/Right to Explanation*

Article 22 GDPR is intended to protect the individual from being made subject to decisions based solely on an automated assessment and evaluation of the subject's personal profile, because this would risk degrading the individual to a mere object of computer-assisted programs. Against this background, the GDPR imposes additional obligations to provide information in situations where the responsible controller utilises automated decision-making procedures in Articles 13(2)(f), 14(2)(g), and 15(1)(h) GDPR. Pursuant to these provisions, the controller has to provide 'meaningful information about the logic involved' in the data processing.

[15] On artificial neuronal networks see for example Y LeCun, Y Bengio and G Hinton, 'Deep Learning' (2017) *Nature Deep Review* 436 (437); T Sejnowski, *The Deep Learning Revolution* (2018) 37 *et seq*.

[16] A Deeks, 'The Judicial Demand for Explainable Artificial Intelligence' (2019) *Colombia Law Review* 1892 (1833 *et seq*.).

[17] Cf. Recital 32. For this see the following Section II 6(a).

This obligation may be called into question[18] when considering the aforementioned difficulties that controllers may face when tasked with providing information about complex and potentially inexplicable (autonomous) AI processes and the results based thereon. In these scenarios, the controller should merely have to provide (and the subject should merely be entitled to) general information on the functioning of the specific AI technology, whereas a right to a substantiated explanation should be rejected. In accordance with Article 35(3)(a) GDPR, an evaluation of personal data which is based on automated processing requires a data protection impact assessment. It should also be emphasised that the use of AI as such is not restricted as of today. Instead, the restrictions apply solely to decision-making processes based on the use of AI.

3. *Purpose Limitation/Change of Purpose*

Pursuant to the principle of purpose limitation as set out in Article 5(1)(b) GDPR, the purposes for processing and collection of (personal) data must be specified and made available to the data subject in a concise and intelligible way.[19] This principle also applies to any further processing of data. The requirement of a pre-defined purpose limitation generally opposes the basic concept of AI, according to which AI should develop independently (or possibly within a certain pre-defined framework) and should be used for purposes not defined in advance.[20] Against this backdrop, the prescription of purpose limitations threatens to impede the (unhindered) development and potentials of AI technologies.[21] Thus, the limitation of legitimate purposes of data processing may lead to a considerable restriction of technological AI potentials.[22] In situations in which AI can (and frequently even should) lead to unforeseen and possibly unforeseeable applications and results, it can, therefore, be very challenging to find an appropriate equilibrium between the principle of purpose limitation and the innovation of AI technologies. In many AI scenarios, it is virtually impossible to predict what the algorithm will learn. Furthermore, the purpose in the sense of Article 5(1)(b) GDPR may change in the course of the (autonomous) development of self-learning AI, especially as the relevant objectives of the data processing may not be known at the time of data collection. Moreover, it is reasonable to be concerned about a distortion of the results (freely) generated by AI tools as potentially induced by data protection law, if such technologies are only granted restricted (or no) access to certain data sources.

[18] In favour of such a right to explanation B Goodman and S Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2017) 38(3) *AI Magazine* 50, 55 *et seq*.; in contrast S Wachter, B Mittelstadt, and L Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the GDPR' (2017) 7 (2) *IDPL* 76.; cf. also M Temme, 'Algorithms and Transparency in View of the New GDPR' (2017) 3(4) *EDPL* 473, 481 *et seq*.; *L Edwards* and *M Veale*, 'Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For' (2017) 16 *DLTR* 18; critical of the GDPR's significance in principle for AI methods also R van den Hoven van Genderen, 'Privacy and Data Protection in the Age of Pervasive Technologies in AI and Robotics' (2017) 3(3) *EDPL* 338, 346 *et seq*.; on the ethical dimension and the efforts to supplement Convention No 108 of the Council of Europe with corresponding transparency provisions, see Committee of Experts on Internet Intermediaries, *Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques* (DGI (2017)12) 13 *et seq*. in particular Algorithms and Possible Regulatory Implications.

[19] See the above comments on transparency Sub II 1.

[20] Mitrou, 'Data Protection' (n 14) 20; N Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10(1) *Law, Innovation and Technology* 40, 56 (hereafter Purtova, 'The Law of Everything').

[21] N Wallace, and D Castro, *The Impact of the EU's New Data Protection Regulation on AI*, 14 (Centre for Data Innovation Policy Brief, 2018) https://euagenda.eu/upload/publications/untitled-140069-ea.pdf (hereafter Wallace and Castro, 'Data Protection Regulation').

[22] Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*, 18 (Datatilsynet Report, 2018) www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

There is, thus, a notable risk of conflicts between the interests and objectives of the individual on the one hand and public welfare on the other. In order to avoid such conflicts, it is crucial to explicitly list the application and use of AI as one of the purposes for the collection of data. Data controllers should, therefore, seek to identify, document, and specify the purposes of future data processing at an early stage. Where these measures are not taken, the requirements for a permissible change of purpose follow from Article 6(4) GDPR.

Article 6(4) GDPR, which addresses purpose changes, lists a number of criteria for the evaluation of the compatibility of such changes in situations where the data processing is carried out for purposes other than the ones for which the data has been originally collected. This creates a direct link to the principle of purpose limitation as laid down in Article 5(1)(b) GDPR. It should further be emphasised that the compatibility of a change of purpose with the original purpose does not affect the cumulative prerequisites for the lawfulness of the processing in question. Because Article 6(4) GDPR itself does not constitute a legal basis for the processing of data for other purposes, recourse must be taken to Article 6(1) subpara. (1) GDPR which requires the existence of a legal justification also for other purposes. In consequence, the controller is responsible to ensure that the data processing for the new purpose is compatible with the original purpose and based upon a legal justification in the sense of Article 6(1) subpara. (1) GDPR. In many cases, relevant personal data will not have been collected for the purposes of training or applying AI technology.[23] In addition, controllers may sometimes have the hope or expectation to subsequently use the collected data for other purposes, for instance in exploratory data analyses. If one were to pursue a more restrictive line of interpretation regarding the change of purposes by applying the standard of Article 6(4) GDPR, it would be impossible to use AI with a sufficient degree of legal certainty. Especially, situations, in which data is generated in different contexts and subsequently combined or used for (new) purposes, are particularly prone to conflict.[24] In fact, this scenario demonstrates the far-reaching implications of and issues arising in connection with the principle of purpose limitation and AI scenarios: if the purpose for the processing of data cannot (yet) be determined, the assessment of its necessity becomes largely meaningless. Where the purpose limitation remains vague and unspecified, substantial effects of this limitation remain unlikely.

4. *Data Minimisation/Storage Limitation*

Pursuant to the principle of data minimisation,[25] personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed. The principle of data minimisation is specified by the requirement of storage limitation (as will be elaborated in the following) and the provisions concerning data protection through the implementation of technical measures and data protection 'by design and by default'.[26] Similarly to the principles described above, the principle of data minimisation oftentimes directly contradicts the general concept of AI technologies which is based on and requires the collection of large

[23] On the consequences of the prohibition on repurposing data see Wallace and Castro, 'Data Protection Regulation' (n 21) 14.

[24] M Butterworth, 'The ICO and Artificial Intelligence: The Role of Fairness in the GDPR Framework' (2018) 34(2) *Computer Law & Security Review: The International Journal of Technology Law and Practice* 257, 260 (hereafter Butterworth, 'GDPR Framework').

[25] Cf. GDPR, Article 5(1)(c).

[26] GDPR, Article 25.

amounts of data.[27] Given the very nature of AI applications, it is exceedingly difficult to make any kind of prediction regarding the type and amount of data necessary in constellations which have yet to be determined by the application itself. In addition, the notion of precautionary protection of fundamental rights by way of data avoidance openly conflicts with the high demand for data in any given AI scenario.[28]

The principle of storage limitation[29] prescribes that where personal data is stored, the identification of the data subject is only permissible for as long as this is necessary for the processing purposes. This principle also poses considerable difficulties in AI constellations, because the deletion or restriction of personal data after the fulfilment of their purpose can significantly impede both the development and use of AI technologies. According to Recital 39 sentences 8 and 10 of the GDPR, the period for which personal data is stored must be limited to a strict minimum. The controller should further establish time limits for the data's erasure or their periodic review. Correspondingly, Article 17 GDPR contains the data subject's right to demand the immediate erasure of any data concerning him or her under certain conditions.[30]

5. *Accuracy/Integrity and Confidentiality*

Another principle of data protection law which may be affected in AI scenarios is the principle of accuracy as set out in Article 5(1)(d) GDPR. This principle is intended to ensure that the collected (personal) data accurately depicts reality so that the affected data subjects will not suffer any disadvantages resulting from the use of inaccurate data. In situations in which the procedure and systems used for the processing of data present themselves as a 'black box' to both data subject and controller, it can be very difficult to detect inaccurate information and to restore their accuracy.[31] However, situations concerning the accuracy of data require a distinction between data input and output; as the latter is a result of data-processing analyses and processes – also and in particular in situations involving AI – it will regularly constitute a (mere) prognosis.

Pursuant to Article 5(1)(f) GDPR, personal data must be processed in a manner that ensures their appropriate security. The controller is thereby required to take adequate measures to ensure the data's protection against unauthorised or unlawful processing and against accidental loss, destruction, or damage.

6. *Lawfulness/Fairness*

The lawfulness of data processing[32] requires a legal basis authorising the processing of data as the normative concept of data protection law envisages a prohibition subject to authorisation. In order to be deemed lawful in the sense of Article 5(1)(a) GDPR, the processing must fulfil at least one of the prerequisites enumerated in Article 6(1) GDPR. In this context, Article 6(1) subpara. 1(b) GDPR permits the processing of data if it is necessary for the performance of a contract which the data subject is party to or for the implementation of pre-contractual measures.

[27] Butterworth, 'GDPR Framework' (n 24) 260.
[28] T Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) 47 *Seton Hall Law Review* 995, 1005 *et seq*.
[29] GDPR, Article 5(1)(e).
[30] On Article 17 and the implications for AI technologies see M Humerick, 'Taking AI Personally: How the EU Must Learn to Balance the Interests of Personal Data Privacy & Artificial Intelligence' (2018) 34 *Santa Clara High Tech L.J.* 393, 407 *et seq*.
[31] On AI and the accuracy principle see Butterworth, 'GDPR Framework' (n 24) 257, 260 *et seq*.; Mitrou, 'Data Protection' (n 14) 51 *et seq*.
[32] GDPR, Article 5(1)(a) alt. 1 and 2.

However, in scenarios involving AI, such pre-contractual constellations will not arise regularly. Similarly, AI scenarios are very unlikely to fall within the scope of any of the other authorisations listed in Article 6(1) subpara. (1) GDPR which include the existence of a legal obligation, the protection of vital interests, or the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.[33]

In contrast, the authorisations set out in Article 6(1) subpara. (1)(a) and (f) GDPR, which are based on the data subjects' consent[34] and the balancing of interests[35] are of great practical importance for the development and use of AI applications. Especially in connection with authorisations relying on consent, attention must be paid to the data subject's right to withdraw his or her consent[36] and to provisions regulating the processing of special categories of personal data.[37]

a. Consent

The most prominent justification for the processing of data is the subject's consent.[38] The requirements for consent can be derived from a conjunction of the provisions stipulated in Article 4(11), Article 6(1) subpara. 1(a), Article 7, and Article 8 GDPR as well as from the general principles of data protection law. The processing of data is only lawful to the extent that consent has been given, meaning that the data subject must give his or her consent for one or more specific purpose(s).[39] Thus, the scope of the justification is determined by the extent of consent. It should also be pointed out that abstract purposes such as 'advertisement' or 'IT-security' are insufficient.[40] This will also apply in the context of AI. Furthermore, Article 4(11) defines consent as the 'freely given' and 'informed' indication of the subject's declaration of intent. The requirement of an 'informed' decision corresponds directly with the previously elaborated principle of transparency[41] which is also laid down in Article 7(2) GDPR. In an AI scenario, this requirement gives rise to further tension between the controllers' obligation to provide adequate information on the one hand and the information's comprehensibility for the average data subject on the other.

The requirement of 'specific' and 'informed' consent may also pose significant challenges where the controller neither knows nor is able to foresee how and for which purposes the personal data will be processed by self-learning and autonomous AI systems. In principle, the practicability of a consent-based justification may be called into question, particularly when considering the voluntary element of such consent in situations lacking any viable alternatives or scenarios of market dominance. In this regard, it may be said that the requirements of a justification based on consent are more fictional than practicable, especially in view of the ubiquity of data-related consent agreements: 'no one has ever read a privacy notice who was not paid to do so.'[42]

b. Withdrawal of Consent

In addition to the fulfilment of the requirements for a consent-based justification, the technical and legal implementation of the withdrawal of consent as set out in Article 7(3) sentence 1 GDPR

[33] GDPR, Article 6(1) subpara (1)(c)–(e).
[34] GDPR, Article 6(1) subpara (1)(a).
[35] GDPR, Article 6(1) subpara (1)(f).
[36] GDPR, Article 7(3).
[37] GDPR, Article 9.
[38] GDPR, Article 6(1) subpara (1)(a).
[39] For information on earmarking see Section II 3.
[40] Cf. Article 29 Data Protection Working Party, 'Guidelines on Consent under Regulation 2016/679' WP 259 rev. 01, 10.
[41] For transparency see the Section II 1.
[42] Butterworth, 'GDPR Framework' (n 24) 257, 262 *et seq*.

is also highly problematic. According to this provision, the data subject has the right to withdraw his or her consent at any time and without having to adhere to any formal requirements. After the consent has been effectively withdrawn, the justification for the processing of data in the sense of Article 6(1) subpara. 1(a) GDPR ceases to exist. In consequence, any further processing of data will only be lawful if, as a substitute, another ground for justification were to apply.[43] Furthermore, a distinction must be made between the right to withdraw in the aforementioned sense, the right to object to unconsented processing of data as regulated by Article 21 GDPR, and, finally, a generally permissible time limitation. As a consequence of the withdrawal of consent, the controller is required to erase the relevant personal data. In cases involving the use of AI, especially scenarios in which certain data is used to train an AI application, it is doubtful whether (and if so, to what extent) the imposition of an obligation to delete is even practicable.[44]

c. Balancing of Interests

The justification based on a balancing of interests allows the processing of personal data in cases where there cumulatively exists (i) a legitimate interest pursued by the controller or by a third party and, (ii) where the processing is necessary to safeguard these legitimate interests, and (iii) where these interests are not overridden by the interests or fundamental rights and freedoms of the data subject who requires the protection of his or her data. The vague wording of this provision is likely to give rise to complications, which do not only apply in the context of AI. For instance, the GDPR does not provide any specific points of reference regarding the general admissibility of and the specific requirements for the processing of data in connection with the balancing of interests within the meaning of Article 6(1) subpara. 1(f) GDPR.

Thus, the task to specify the requirements of the abovementioned balancing process is mostly assigned to academic discourse, courts, and public authorities. However, such an interpretation of the GDPR must, in any case, comply with and adhere to the objective of a consistent standard of (data) protection throughout the EU.[45] It is, therefore, subject to the requirement of a harmonised interpretation of the law which, in turn, is intended to guarantee equal data processing conditions for all market participants in the EU.[46] In addition, by establishing codes of conduct designed to contribute to the appropriate application of the GDPR, Member States are encouraged to provide legal certainty by stating which (industry-specific) interests can be classified as legitimate in the sense of Article 6(1) GDPR. Finally, the European Data Protection Board may, pursuant to Article 70(1)(e) GDPR, further ensure the consistent application of the Regulation's provisions by issuing guidelines, recommendations, and best practices, particularly regarding the practical implementation of the aforementioned balancing process.

d. Special Categories of Personal Data

Article 9 GDPR establishes a separate regulatory regime for special categories of personal data and prohibits the processing of these types of data. These include, for instance, genetic and biometric data, or data concerning health, unless their processing falls under one of the exemptions listed in Article 9(2) GDPR. In accordance with Article 22(4) GDPR, automated decisions, including profiling, must not be based on sensitive data unless these exemptions

[43] It has to be taken into account that it could present itself as contradictory behaviour if, in the case of the omission of consent, an alternative legal justification is applied.
[44] Wallace and Castro, 'Data Protection Regulation' (n 21) 12 *et seq*.
[45] GDPR, Recital 13.
[46] GDPR, Recitals 9 and 10.

apply. Furthermore, the processing of large amounts of sensitive data, as referred to in Article 35 (3)(b) GDPR, requires an obligatory data protection impact assessment. Overall, the use and application of AI impose new challenges for the protection of sensitive data. The accumulation of personal data in conjunction with improved methods of analysis and (re-)combination will certainly increase the likelihood of cases affecting potentially sensitive data within the meaning of Article 9 and Recital 51 of the GDPR. Consequently, an increasing amount of data may fall under the prohibition of Article 9(1) GDPR. It is, therefore, necessary to closely follow new trends and developments in the technical field, including but not limited to AI, in order to correctly determine the scope of application of Article 9 GDPR. These findings leave controllers with considerable (legal) uncertainties regarding their obligations.

In light of the new possibilities for a fast and effective AI-based evaluation of increasingly large amounts of data (i.e. big data), the question arises whether metadata, source data, or any other types of information which, by themselves, generally do not allow the average observer to draw any conclusions as to the categories mentioned in Article 9(1) GDPR, nevertheless fall under this provision. If so, one may consider adding the application of AI technology to the list of potential exemptions under Article 9(2) GDPR. In this context, however, regard must be paid to the principle of purpose limitation as previously mentioned.

7. *Intermediate Conclusion*

Given its rather broad, oftentimes undefined and vague legal terminology, the GDPR, in many respects, allows for a flexible application of the law. However, this flexibility goes hand in hand with various (legal) uncertainties. These uncertainties are further perpetuated by the GDPR's notable and worrisome lack of reference to and regulation of AI-specific constellations. As shown above, these constellations are particularly prone to come into conflict with the general principles of data protection as set out in Article 5(1) GDPR and as specified and reiterated in a number of other provisions. In this context, the principles of data minimisation and storage limitation are particularly problematic. Other conflicts, especially involving the GDPR's principles of purpose limitation and transparency, may arise when considering the rather complex and ambiguous purposes and structures for the processing of data as well as the open-ended explorative analyses frequently observed in AI-scenarios. This particularly applies to subsequent changes of purpose.[47] It must also be emphasised that the requirement of transparency serves as a regulatory instrument to ensure the lawfulness of data processing and to detect tendencies of dominance[48] or, rather, the abuse thereof. However, legal uncertainties entail considerable risks and burdens for controllers implementing AI technologies which are amplified and intensified by the GDPR's new and much stricter sanctions regime.[49] Finally, it has to be pointed out that these conflicts by no means only apply to known concerns of data protection law, but rather constitute the starting point for new fundamental questions in this field.

III. COMPLIANCE STRATEGIES (*DE LEGE LATA*)

Based on these findings, it is necessary to examine potential strategies to comply with the provisions of the GDPR and to establish a workable and resilient framework which is capable of fostering the future development and application of AI technologies under the given legal framework. It should

[47] Cf. GDPR, Article 6(4).
[48] For this see the following Section IV 3.
[49] Wallace and Castro, 'Data Protection Regulation' (n 21) 18 *et seq*.

also be emphasised that the enactment of the GDPR has fundamentally increased the requirements for compliance with data protection law. This development was further accompanied by substantially higher sanctions for the infringement of data protection law.[50] In addition to potential sanctions, any infringement of data protection law may also give rise to private damage claims pursuant to Article 82 GDPR which cover both material and non-material damage suffered by the data subject. The legally compliant implementation of AI may further be impeded by the interplay and collision of different or conflicting data protection guarantees. Such guarantees can, for instance, be based on data protection law itself, on other personal rights, or on economic and public interests and objectives. In an AI context, this will become particularly relevant in connection with the balancing of interests required by Article 6(1) subpara. 1(f) GDPR.

Article 25 GDPR contains the decisive normative starting point for data protection compliance, in other words the requirement that data protection-friendly technical designs and default settings must be used. However, the rather vague wording of this provision (again) calls for an interpretation as well as specification of its content. The obligation of the responsible controller to implement appropriate technical and organisational measures is essential in terms of data protection compliance. Overall, the GDPR pursues a risk-based approach.[51] From a technical and organisational point of view, it is, thus, necessary to ask how the protection of personal data can be achieved by way of a data protection management system and other measures, for instance through anonymisation and pseudonymisation. The starting point of these considerations is the connection between the data in question and an individual (personal reference), which is decisive for the opening of the substantive scope of application of the GDPR.[52]

1. *Personal Reference*

The existence of such a personal reference is a necessary prerequisite for the application of the GDPR. From a factual point of view, as set out by Article 2(1) GDPR, the GDPR applies in cases of a 'wholly or partially automated processing of personal data and for non-automated processing of personal data stored or to be stored in a file system'. Therefore, it must be asked whether, in a given case and under specific circumstances, personal data is being processed.[53] According to the legal definition stipulated in Article 4(1) GDPR, personal data is

> any information relating to an identified or identifiable natural person [...]; an identifiable natural person is one who can be identified directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or one or more specific characteristics expressing the physical, physiological, genetic, psychological, economic, cultural or social identity of that natural person.

According to the pertinent case law of the European Court of Justice (ECJ), it is sufficient for the responsible data controller to have legal means at his or her disposal to make the data of the third party available (so-called absolute personal reference); this can also encompass detours via state authorities.[54]

[50] Cf. Article 83 GDPR: up to € 20 million or 4% of worldwide turnover.
[51] Cf. GDPR Articles 30(5), 33(1), 35, 36, and 37(1).
[52] Already critical about the old legal situation before the GDPR regarding the (legal) uncertainties regarding personal references and anonymisation J Kühling and M Klar, 'Unsicherheitsfaktor Datenschutzrecht – Das Beispiel des Personenbezugs und der Anonymität' (2013) *NJW* 3611.
[53] On the expanding scope of personal data under the GDPR see Purtova, 'The Law of Everything' (n 20) 40, 43 *et seq*.
[54] CJEU, C-582/14, *Patrick Breyer v Bundesrepublik Deutschland* (19 October 2016), paras 47 *et seq*.

The application of the GDPR – and thus the application of its strict regulatory regime – could be avoided by way of, for instance, the data's anonymisation. Article 3(2) of Regulation No. 2018/1807 concerning the free movement of non-personal data states that, in the event of personal and non-personal data[55] being inseparable, both sets of rules (regarding personal and non-personal data) must, in principle, be applied. However, in many cases, it will not be easy to determine with any (legal) certainty whether and to what extent data records may also contain personal data. Hence, in order to remain on the 'safe side' regarding the compliance with the current data protection regime, controllers may feel the need to always (also) adhere to the provisions and requirements of the GDPR even in cases where its application may be unnecessary. This approach may result in considerable (and needless) expenditures in terms of personnel, material, and financial resources.

a. Anonymisation

In contrast to personal information in the aforementioned sense, the GDPR does not apply to anonymous information because they are, by their nature, the very opposite of personal. Recital 26 of the GDPR states: 'The data protection principles should therefore not apply to anonymous information, i.e. information which does not relate to an identified or identifiable natural person, or personal data which has been anonymised in such a way that the data subject cannot or can no longer be identified.'

The Regulation, therefore, does not address the processing of such anonymous data, including data for statistical or research purposes. It follows from the aforementioned Recital that, when it comes to the identifiability of an individual person, the technological capabilities and developments available at the time of the processing must be taken into account. However, when it comes to the technical specifications with regard to the actual anonymisation process, the GDPR, with good reason, does not stipulate a specific procedure to follow. This lack of a prescribed procedure not least benefits innovation and development of new technologies and the concept of technological neutrality.[56] The relevant time of evaluation is always the time of the processing in question.

This is further not changed by mere reference to the fact that almost all anonymised data may be restored by means of advanced sample formations, because such an objection is far too broad and, thus, certainly falls short of the mark.[57] Nevertheless, it should also be noted that, with respect to data relating to location, an anonymisation is considered virtually impossible. Thus, Article 9 GDPR bears particular significance when it comes to the inclusion of location data in the relevant applications. In any case, the issue of de-anonymisation, for which especially the available data stock, background knowledge, and specific evaluation purposes have to be considered, remains highly problematic. According to Recital 26 of the GDPR, in order to identify means likely to be used for the identification of an individual, all objective factors such as costs, time, available technologies, and technological developments, should be considered. In this context, the continuously more advanced big data analysis techniques tend to lead to an ever further reaching re-identification of persons in a constantly growing data pool. In addition, the change of the underlying technological framework and the conditions thereof may (over time) result in the 'erosion' of the former anonymisation and subsequently uncover or expose a

[55] Regulation EU 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free movement of non-personal data in the European Union (2018) OJ L 303, 59.

[56] Due to legal uncertainties companies might be deferred from using such data, Wallace and Castro, 'Data Protection Regulation' (n 21) 15.

[57] On the discussion see Purtova, 'The Law of Everything' (n 20) 40, 42 *et seq*.

personal reference of the respective data. Naturally, the consequential (legal) uncertainties may pose a considerable risk and problem to users and other affected parties, especially with respect to issues of practical manageability and incentive. In this context, the opinion on Anonymisation Techniques of the Article 29 Data Protection Working Party (particularly relating to the robustness of randomisation and generalisation) may give helpful indications, but will certainly not be the solution to all potential issues arising in connection thereto.[58] Thus, the findings and developments mentioned earlier give rise to well-founded doubts as to whether the comprehensive anonymisation of data can be successfully achieved under the current framework conditions (e.g. technological progress and available data volumes).

b. Pseudonymisation

According to the legal definition provided in Article 4(5) GDPR, pseudonymisation means the processing of personal data in such a way that personal data cannot (any longer) be assigned to a specific data subject without the use of additional information. Although the GDPR does not expressly permit or privilege the processing of personal data in the event of a pseudonymisation, there are a number of substantial incentives to carry out such a pseudonymisation: in the case of a pseudonymisation, the balancing of interests within the meaning of Article 6(1) subpara. 1(f) GDPR is more likely to sway in favour of the processor. Furthermore, in the case of data protection violations pursuant to Article 34(3)(a) GDPR, the obligation to notify the data subject does not apply in cases of encryption as a sub-category of pseudonymisation. In addition, the procedure may decrease the need for further technical and organisational protection and may, in the event of a previously mentioned change of purpose, be included as a factor in the balancing process as required by Article 6(4)(e) GDPR. Pseudonymisation, therefore, has the potential to withdraw the processing of certain data from the scope of the GDPR and to avoid the application of the Regulation's strict requirements.

c. Synthetic Data

Another possibility to avoid a personal reference and, thus, the application of the GDPR is the production and use of synthesised data which constitute a mere virtual representation of the original set of data. The legal classification of synthetic data is directly linked to the existence or producibility of a personal reference. As a result, the lack of a personal reference allows synthetic data to be equated to anonymous data. In this context and in connection with all related questions, the decisive issue is, again, the possibility of a re-identification of the data subject(s). Another potential problem that must be taken into account relates to eventual repercussions on the data subjects of the (underlying) original data set from which the synthetic data were generated. For instance, processing operations subject to the provisions of the GDPR may hereby arise due to the predictability of sensitive characteristics resulting from a combination of multiple data sets.

2. *Responsibility*

The question of who is responsible for the compliance with the requirements of data protection and to whom data subjects can turn in order to exercise their rights is of great importance.[59] Article 4(7) GDPR defines the data controller as 'the natural or legal person, public authority,

[58] See Article 29 Data Protection Working Party, 'Opinion 5/2014 on Anonymisation Techniques' WP 216.
[59] In detail see Mitrou, 'Data Protection' (n 14) 60 *et seq*.

agency or other body which alone or jointly with others determines the purposes and means of the processing of personal data'. In practice, an essential (distinguishing) characteristic of a data controller within the sense of the GDPR is, thus, the authority to make a decision about the purposes and means of data processing. In a number of recent rulings, the ECJ has further elaborated the criterion of responsibility by specifying the nature and extent of a controller's decision concerning the purpose and means of processing personal data: Facebook Fanpage,[60] Jehovah´s Witnesses,[61] and Fashion ID/Facebook Like Button.[62] According to the (previous) case law of the ECJ, those involved in the processing of data do not necessarily have to bear an equal amount of responsibility. Instead, the criterion of responsibility is met if the participants engage in the data processing at different stages and to varying extents, provided that each participant pursues its own purposes for the processing.

3. *Privacy by Default/Privacy by Design*

Article 25 GDPR contains provisions concerning the protection of data by way of (technology) design and data protection-friendly default settings.[63] The first paragraph of the provision stipulates the principles for privacy by design, that is, the obligation to design technology in a manner that facilitates and enables effective data protection (in particular to safeguard the implementation of data-protection principles such as data minimisation). In its scope, the provision is limited to an enumeration of various criteria to be taken into account by the controller with regards to the determination of appropriate measures and their respective durations. The provision does not further specify any concrete measures to be taken by the responsible controller – with the exception of pseudonymisation as discussed earlier. In addition, Article 25(2) GDPR sets out the principle of privacy by default, in other words, the controller's obligation to select data protection-friendly default settings to ensure that only data required for the specific purpose are processed. Finally, in order to demonstrate compliance with the requirements of Article 25(1) and (2) GDPR, Article 25(3) allows the use of an approved certification procedure in accordance with Article 42 GDPR. The challenges previously described typically arise in cases where the GDPR's transparency requirements coincide with complex AI issues, which – by themselves – already present difficulties for the parties concerned. Against this background, certification procedures, data protection seals, and test marks in the sense of Article 42 GDPR could represent valuable instruments on the way to data protection compliance.

4. *Data Protection Impact Assessment*

The data protection impact assessment pursuant to Article 35 GDPR addresses particularly high-risk data processing operations with regard to the rights and freedoms of natural persons. The provision requires the controller to carry out a preventive review of the potential consequences

[60] ECJ, Case C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH (*Facebook Fanpage Case*, 5 June 2018).

[61] ECJ, Case C-25/17 *Tietosuojavaltuutettu v Jehovan todistajat*-uskonnollinen yhdyskunta (*Jehovah's Witnesses Case*, 10 July 2018).

[62] ECJ, Case C-40/17 *Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW eV* (29 July 2019).

[63] For detail on privacy by default and privacy by design see L Bygrave, 'Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision Making' in K Yeung and M Lodge (eds), *Algorithmic Regulation* (2019) 9 *et seq*. https://ssrn.com/abstract=3329868.

of any processing operations likely to result in such a high risk and to subsequently select and implement the appropriate risk-minimising remedial measures. The obligation to carry out a data protection impact assessment serves the purpose to ensure the protection of personal data and, thus, the compliance with the provisions of the Regulation.[64] At the same time, Articles 35 (4) and (5) require the responsible supervisory authority to establish and make public a list of the kind of the specific processing operations which require such an impact assessment[65] and of those operations which do not require an assessment.[66] These lists are intended to ensure legal certainty for and equal treatment of responsible controllers and to facilitate transparency of all parties concerned. By including processing operations, for which a data protection impact assessment must be carried out, in a list,[67] the supervisory authority can also positively establish an obligation to carry out a data protection impact assessment.

Furthermore, it should be noted that Article 35(1) GDPR explicitly requires the conduction of a data protection impact assessment 'in particular' where 'new technologies' are used. Naturally, this provision is of particular relevance in cases where large amounts of data are processed using 'new' AI systems and technologies and might indicate that the use of such applications may automatically trigger the need for a comprehensive and onerous impact assessment. The GDPR does not explicitly provide any examples of technologies or areas of technology which qualify as 'new'. However, a new technology is likely to pose a high risk to the rights and freedoms of natural persons if it enables the execution of large-scale processing operations which allow the processing of large quantities of personal data at a regional, national, or supranational level and which may involve data relating to a large number of individuals and data of a particularly sensitive nature. Thus, developments such as Smart Car, Smart Health, Big Data, and Tracking procedures as well as new security and monitoring technologies are likely to fall within the scope of Article 35(1) GDPR, hence requiring controllers using and offering such technologies to conduct a data protection impact assessment in accordance with Article 35 GDPR.[68] In the context of AI systems, it remains highly doubtful whether such an obligation would even be feasible, given the fact that self-learning programs develop continuously and more or less unpredictably.

5. *Self-Regulation*

In order to specify, construe, and interpret the large number of indeterminate legal provisions of the GDPR, it is also necessary to give consideration to the elements of self-regulation or co-regulation.[69] Article 40 GDPR gives associations and other bodies the possibility to draw up, amend, or extend rules of conduct which clarify the application of the GDPR. Thus, pertinent rules of conduct can be developed (e.g. by means of best practices) which can subsequently be approved by the responsible supervisory authorities[70] or given general binding force by the EU Commission. In addition, certification procedures, data protection seals, and test marks[71] could also serve as valuable instruments when it comes to the compliance with data protection law.

[64] Purtova, 'The Law of Everything' (n 20) 77.
[65] GDPR, Article 35(4).
[66] GDPR, Article 35(5).
[67] Cf. GDPR Article 35(4).
[68] Mitrou, 'Data Protection' (n 14) 65 *et seq*.
[69] Cf. GDPR Articles 40–43.
[70] GDPR Article 40(5).
[71] Cf. GDPR Article 42(1).

On the basis of Article 42 GDPR, data controllers may also voluntarily seek certification of their data processing operations by the responsible supervisory authority or an accredited body within the meaning of Article 43 GDPR. Recital 100 of the GDPR emphasises that the associated certification procedures, data protection seals, and marks are intended to increase transparency and improve compliance with the GDPR's requirements.

IV. LEGAL POLICY PERSPECTIVES (*DE LEGE FERENDA*)

In view of the earlier points, it becomes evident that the use and implementation of AI-based technologies necessitates a thorough review of the current data protection framework. Such a review may indicate the need for the modification, amendment, or development of the GDPR's current regime. From a legal policy perspective, legislative initiatives should hereby be the main point of focus.

1. *Substantive-Law Considerations*

From a substantive-law point of view, one key element of the GDPR that merits a closer examination is the personal reference as a prerequisite for the GDPR's application. This is not least due to the structural narrowness of the personal reference in its current definition as well as its frequent lack of adequate relevance. Presently, the personal reference as a connecting criterion only insufficiently reflects the existing multiple rationalities of data processing constellations and lacks the capability to take into account the specific characteristics of each case-by-case context. In fact, the one-size-fits-all-approach of the GDPR does not appropriately distinguish between different risk situations, which means that – due to the ubiquitous relevance of personal data – there exists the risk of an excessive application of the law. Among others, this certainly applies to issues relating to the use of AI as presently discussed. With this in mind, it is both necessary and important to create sector-specific regulations for AI constellations, for instance regarding the permissibility of data processing and the specific requirements thereof. Furthermore, the ubiquity of data processing operations in the present age of digitalisation frequently calls into question the general concept of data protection in its current state. It is, therefore, necessary to (at least partially) move away from the current approach, in other words, the prohibition subject to permission in favour of a more general clause. Such a provision should differentiate between different data protection requirements according to specific risks that specific situations are likely to pose. Such a stringent risk-based approach would have the advantage of facilitating the weighing and balancing of the interests of all affected parties as well as appropriately taking into account their respective purposes for protection. In addition, the readjustment of the objectives that the GDPR serves to protect may help to realise an adequate protection of an individual's personality and privacy rights whilst also incentivising the development and use of AI applications. In this context, the overarching objective should always be to reassess the balance of interests pursued by data subjects, responsible processors, third parties, and the public welfare in general.

Another issue that ought to be addressed relates to the granting of access to data and the corresponding rights of usage. This further encompasses questions as to the law of obligations in a data law context, data ownership, and data economics. Finally, due consideration should be given to whether the existing legal framework should be supplemented by specific provisions governing the use of AI. These provisions should not least be capable of overcoming the currently existing tensions resulting from the bi-dimensional, two-person relationship between

controller and data subject. This could necessitate an amendment of data protection law with regard to AI in order to move away from an approach based solely on the individual and to appropriately take into account the challenges that may arise in connection with the quantity, heterogeneity, inter-connectivity, and dynamism of the data involved. Such an amendment should be accompanied by more systematic protective measures. A valuable contribution could hereby be made by technical design and standardisation requirements. In addition, all of these measures must be safeguarded and supported by way of an adequate and effective supervisory and judicial protection.

2. *Conflicts between Data Protection Jurisdictions*

Due to the cross-border ubiquity of data (processing) and the outstanding importance of AI-related issues, efforts must be made to achieve a higher degree of legal harmonisation. Ideally, such a development could result in the establishment of an overarching supra- or transnational legal framework, containing an independent regulatory regime suited to the characteristics of AI. Such a regime would also have to take into account the challenges resulting from the interplay of multi-level legal systems as well as the conflicts arising between different data protection legal regimes. For instance, conflicts may arise when the harm-based approach of US data protection law, which is focused on effects and impairments, the Chinese system, which allows for far-reaching data processing and surveillance (e.g. a Social Credit System), and the GDPR approach, which is based on a preventive prohibition subject to permission, collide. Assuming that a worldwide harmonisation of the law is hardly a realistic option in the foreseeable future, it is important to aim for an appropriate balance within one's 'own' data (protection) regime.

3. *Private Power*

In connection with the transnationalisation of the legal framework for data protection and the conflicts between different regulatory regimes, regard must also be paid to the influence exerted by increasingly powerful private (market) players. This, naturally, raises questions as to the appropriate treatment and, potentially, the adequate containment of private power, the latter of which stems from considerations regarding the prevention of a concentration of power and the sanctioning of the abuse of a dominant market position. However, the GDPR itself does not directly stipulate any specific protective measures governing the containment of private power. Legal instruments capable of addressing the aforementioned issues must, therefore, be found outside of the data protection law body. For this purpose, recourse is frequently taken to the (unional or national) competition law, because it expressly governs questions relating to the abuse of market power by private undertakings and, in addition, provides a reliable system and regulatory framework to address such issues. In this regard, the German Federal Cartel Office (FCO) served as a pioneer when it initiated proceedings against Facebook for the alleged abuse of a dominant market position through the use of general terms and conditions contrary to data protection law, specifically the merging of user data from various sources.[72] In any case, the role and power of private individuals as an influential force in the field of data protection should certainly not be underestimated. In fact, by establishing new technological standards and,

[72] BKartA, *Facebook Inc. i.a. Case – The use of abusive business terms pursuant to Section 19 (1) GWB* (B6–22/16, 6 February 2019).

thereby, elevating their processing paradigms and business models to a *de facto* legal power, they have the potential to act as substitute legislators.

V. SUMMARY AND OUTLOOK

There is an inherent conflict of objectives between the maximisation of data protection and the necessity to make use of (large quantities of) data, which transcends the realms of AI-related constellations. On the one hand, the availability and usability of personal data bears considerable potential for innovation. On the other hand, the possibilities and limitations of data processing for the development and use of AI are (above all) determined by the requirements of the GDPR. In consequence, the permissibility of the processing of personal data must be assessed in accordance and adherence with the powers to collect, store, and process data as granted by the GDPR. The law of data protection thereby imposes strict limits on the processing of personal data without justification or sufficient information of the data subject. These limitations have a particular bearing on issues relating to AI, as it is frequently impossible to make a comprehensive *ex ante* determination of the scope of the processing operations conducted by a self-learning, autonomous system. This is not least due to the fact that such systems may only gain new information and possibilities for application – potentially relating to special categories of personal data – after the processing operation has already started. In addition, the processing of such large amounts of personal data is oftentimes likely to result in a significant interference with the data subjects' fundamental rights. All of these considerations certainly give rise to doubts as to whether a complete anonymisation of data is even a viable possibility under the given framework conditions (i.e. technological progress and available data volumes).

In order to combat these shortcomings of the current data protection framework, the establishment of a separate legal basis governing the permissibility of processing operations using AI-based applications should be considered. Such a separate provision would have to be designed in a predictable, sector-specific manner and would need to adhere to the principle of reasonableness, thus also ensuring the adequate protection of fundamental rights and the rule of law. The GDPR's *de lege lata* approach to the processing of personal data, in other words, the comprehensive prohibition subject to permission leaves controllers – as previously elaborated – in a state of considerable legal uncertainty. As of now, controllers are left with no choice but to seek the users' consent (whereby the requirements of informing the data subject and the need for their voluntary agreement apply restrictively) and/or to balance the interests involved on a case-by-case basis. These input limits not only burden controllers immensely, but are also likely to ultimately limit output significantly, especially in an AI context. In fact, the main principles of the applicable data protection law, (i.e. the principles of transparency, limitation, reservation of permission, and purpose limitation), appear to be in direct conflict with the functioning and underlying mechanisms of AI applications which were, evidently, not considered during the drafting of the GDPR's legal regime. In practice, this is especially problematic considering that the GDPR has significantly increased the sanctions imposed for violations of data protection law.

Multidimensional border dissolutions occur and do mainly affect the levels of technology and law, territories, and protection dimensions: on the one hand, these border dissolutions may promote innovation, but at the same time they threaten to erode the structures of efficient law enforcement. The previously mentioned tensions between the GDPR and the basic concepts underlying AI also raise the fundamental question of whether traditional data protection principles in the age of digitalisation, especially with regard to AI, Big Data, the Internet of Things, social media, blockchain, and other applications, are in need of a review. Among others,

the instrument of consent as a justification in AI constellations, which are typically characterised by unpredictability, and limited explainability, must be called into question. In any case, the legal tools for the protection of privacy need to be readjusted in the context of AI. This also and especially applies to the data protection law regime. Against this background, legislative options for action at national, unional, and international level should be examined. In this context, the protection of legal interests through technology design will be just as important as interdisciplinary cooperation and research.

Overall, (legal) data policy is a central industrial policy challenge that needs to be addressed – not only for AI constellations. Legal uncertainties may cause strategies of evasion and circumvention, which in turn (can) trigger locational disadvantages and enforcement deficits, bureaucratic burdens, and erosion with respect to legal compliance. Thus, AI-specific readjustments of data protection law should – where necessary – prevent imminent disadvantages in terms of location and competition and ensure that technology and law are open to innovation and development. Both new approaches to the interaction between data protection law and AI should be examined and existing frameworks retained (and, where appropriate, further developed). By these means, a modern data and information usage right may be established which does not result in a 'technology restriction right' but rather gives rise to new development opportunities. The legal questions raised and addressed in this article concern not only isolated technical issues but also the social and economic order, social and individual life, research, and science. In this sense, the existing legal framework (the European approach) should be further enhanced/developed to make it an attractive alternative to the approaches taken in the US and China, while the current model of individual protection should be maintained, distinguishing it from the other data protection regimes. With the ongoing GDPR evaluation, it is an opportune time for such an initiative. However, such an initiative requires the cooperation of all actors (users and developers, data protection authorities and bodies, policy and legislation, science, and civil society) in order to reconcile data protection with the openness of technology and law for necessary developments.

18

Data Governance and Trust

Lessons from South Korean Experiences Coping with COVID-19

*Sangchul Park, Yong Lim, and Haksoo Ko**

I. INTRODUCTION

COVID-19 is reshaping history with its unprecedented contagiousness. The epidemic swept the whole world throughout 2020 and beyond. In the case of South Korea (hereafter Korea), the first confirmed case of COVID-19 was reported on 20 January 2020.[1] During the initial phase after the first reported case, the Korean government hesitated to introduce compulsory quarantine for travelers from high-risk countries.[2] It put Korea on a different trajectory compared to other countries which imposed aggressive measures including immigration quarantine from the beginning.[3] The number of confirmed infections increased significantly in a short span of time and, by the end of February 2020, the nation was witnessing an outbreak that was threatening to spiral out of control. Korea appeared to be on the way to becoming the next 'COVID-19 hotspot' after China.[4] Confronting an increasing number of cases of COVID-19, Korea had to weigh among various options for Non-Pharmaceutical Interventions (NPIs). Korea did not take extreme measures such as shelter-in-home and complete lockdowns. Instead, it employed a series of relatively mild measures, including a social distancing order that imposed restrictions on public gatherings and on operating businesses, set at different levels in accordance with the seriousness of the epidemic.[5] A differentiated measure that Korea took was an aggressive contact tracing scheme, which served a complementary role to social distancing.

* This chapter is a revised and expanded version from S Park and Y Lim, 'Harnessing Technology to Tackle COVID-19: Lessons from Korea' (2020) 61 *Inform. Process. [Jōhōshori]* 1025.

[1] Korea Disease Control and Prevention Agency (KDCA), 'A Foreign-Imported Case of Novel Coronavirus Was Confirmed during Immigration Quarantine: The Epidemic Crisis Alert Level Elevated to Warning' (*KDCA*, 20 January 2020) http://ncov.mohw.go.kr/tcmBoardView.do?ncvContSeq=352435&contSeq=352435.

[2] Korea started to impose a compulsory two-week quarantine for travelers from Europe on 22 March, 2020, for travelers from the US on 27 March, 2020, and for travelers from the other countries including China on 1 April, 2020. KDCA, 'COVID-19 Domestic Case Status' (*KDCA*, 27 March 2020) http://ncov.mohw.go.kr/tcmBoardView.do?ncvContSeq=353770&contSeq=353770.

[3] J Summers and others, 'Potential Lessons from the Taiwan and New Zealand Health Responses to the COVID-19 Pandemic' (2020) 4 *Lancet Reg Health West Pac* 10044.

[4] S Park, GJ Choi and H Ko, 'Information Technology-Based Tracing Strategy in Response to COVID-19 in South Korea – Privacy Controversies' (2020) 323(21) *JAMA* 2129.

[5] The central and municipal and/or local governments are authorised to 'restrict or prohibit the aggregation of multiple persons including entertainment, assembly, and rituals' in accordance with Article 49-1(ii) of the Contagious Disease Prevention and Control Act. Based on this provision, the government set the level of social distancing from Level 1 to Level 3 (with the interval being 0.5).

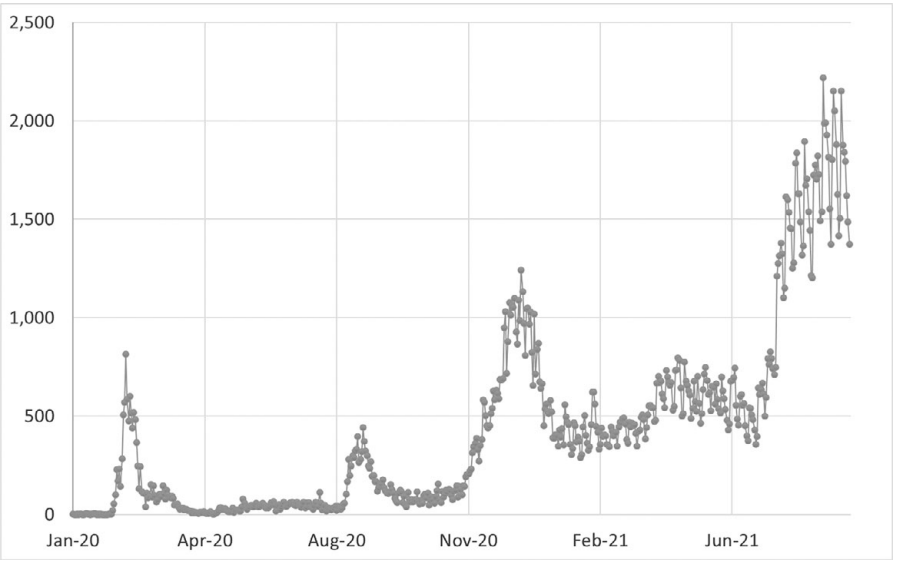

FIGURE 18.1 Daily newly confirmed COVID-19 cases
Note: KDCA, Press Releases (*MOHW*, 20 January 2020 to 1 September 2021), http://ncov.mohw.go.kr

Adopting an effective contact tracing strategy requires, as a pre-requisite, a lawful and technically feasible capability to collect and process relevant personal data including geolocation data. Doing so was possible in Korea because it had already introduced a legal framework for technology-based contact tracing after its bruising encounter with the Middle East Respiratory Syndrome (MERS) in 2015. Based on its previous experience with MERS and the legislative measures and mandates adopted in the course of the MERS outbreak, Korea was well equipped to respond to COVID-19 by swiftly mounting aggressive contact tracing and other data processing schemes when COVID-19 materialised as a significant threat to the public health of its citizens. Thus, the nation's technological infrastructure was mobilized to provide support for epidemiological investigations. The contact tracing scheme, along with a sufficient supply of test kits (such as PCR [polymerase chain reaction] kits for real-time testing) and of personal protective equipment (such as respirators), was perhaps a key contributing factor to Korea's initial success in flattening the curve of infections and deaths, when it had to confront two major outbreaks that occurred around March and August 2020, respectively. Toward the end of 2020, Korea began facing a new round of difficulties in dealing with a third outbreak, and it again actively implemented a contact tracing scheme. As of 00:00, 1 September 2021, the accumulated number of confirmed cases was recorded at 253,445 (0.49% of the total population), including 2,292 total deaths.[6] Figure 18.1 shows the trend of newly confirmed cases.

While the statutory framework introduced after the MERS outbreak provided the necessary means to launch a technology-based response to COVID-19, new challenges arose in the process. In particular, there was an obvious but challenging need to protect the privacy of those infected and of those who were deemed to have been in close contact while, at the same time, maintaining the effectiveness of the responses. This chapter provides an overview of how Korea

[6] KDCA, 'COVID-19 Domestic Case Status (1 September 00:00)' (*KDCA*, 1 September 2021) http://ncov.mohw.go.kr/tcmBoardView.do?brdId=3&brdGubun=31&dataGubun=&ncvContSeq=5878&contSeq=5878&board_id=312&gubun=BDJ.

harnessed the power of technology to confront COVID-19 and discusses some of the issues related to the governance of data and technology that were raised during Korea's experiences.

This chapter is organized as follows: Section II provides an overview of the legal framework which enabled an extensive use of the technology-based contact tracing scheme; Section III explains the structure of the information system that Korea set up and implemented in response to COVID-19; Section IV details the actual use of data for implementing the legal scheme and relevant privacy controversies; Section V further discusses data governance and trust issues; and, finally, Section VI concludes.

II. LEGAL FRAMEWORKS ENABLING EXTENSIVE USE OF TECHNOLOGY-BASED CONTACT TRACING

1. *Consent Principle under Data Protection Laws*

A major hurdle in implementing the pandemic-triggered contact tracing scheme in Korea was the country's stringent data protection regime. Major pillars of the legal regime include the Personal Information Protection Act (PIPA),[7] the Act on Protection and Use of Location Information (LIA),[8] and the Communications Secrecy Protection Act (CSPA).[9] As a means to guarantee the constitutional right to privacy and the right to self-control of personal data, these laws require prior consent from the data subject or a court warrant prior to the collection and processing of personal data, including geolocation data and communications records. Arguably, the consent principle of the Korean law is largely modeled after what can be found in the European Union's (EU's) privacy regime including the General Data Protection Regulation (GDPR). However, Korea's data protection laws tend to be more stringent than the EU's, for instance, by requiring formalities such as the notification of mandatory items when obtaining consent. Certain statutory features of the Korean data protection laws on data collection are as follows.

First, the PIPA is the primary law governing data protection. Under the PIPA, the data subject must, before giving consent to collection, be given notice including the following: (i) the purpose of collection and use, (ii) the items of data collected, (iii) retention and use period, and (iv) (unless data is collected online) the data subject's right to refuse consent and disadvantages, if any, from the refusal.[10] The data subject must, before giving consent to disclosure, be given notice of the recipient and similar items as above.[11] A recent amendment to the PIPA which took place in 2020 allows exceptions to the purpose limitation principle within the scope reasonably related to the purpose for which the personal data is initially collected.[12] The 2020 amendment of the PIPA also grants an exemption to the consent requirement when the processing of pseudonymized personal data is carried out for statistical, scientific research, or archiving purposes.[13] However, these built-in exceptions are not broad enough to cover the processing of personal data for the centralized contact tracing scheme.

[7] Personal Information Protection Act [*Gaein Jeongbo Boho Beop*], Act No 16930 (last amended on 4 February 2020, effective as of 4 February 2020).

[8] Act on Protection and Use of Location Data [*Wichi Jeongboeu Boho Mit Iyong Deung'e Gwanhan Beopryul*], Act No 17689 (last amended on 22 December 2020, effective as of 1 January 2021).

[9] Communications Secrecy Protection Act [*Tongshin Bimil Hobo Beop*], Act No 17831 (last amended and effective on 5 January 2021).

[10] PIPA, Articles 15(2), 39-3(1).

[11] PIPA, Article 17(2).

[12] PIPA, Articles 15(3) and 17(4).

[13] CSPA, Article 28-2(1).

Second, the LIA is a special law that governs the processing of geolocation data such as GPS (global positioning system) data and cell ID. This type of data is usually collected by mobile carriers or mobile operating system operators and is shared with mobile app developers. Under the LIA, a data subject of geolocation data must be given appropriate notice in the standard forms before giving consent to the collection, use, or disclosure of personal geolocation data.[14]

Third, the CSPA governs when and how courts or law enforcers can request communications records including base station data or IP (internet protocol) addresses from carriers or online service providers.[15] Under the CSPA, law enforcers can request data concerning a specific base station (the base station close to the location where the mobile phone user at issue made calls) from mobile carriers in order to deter crime, to detect or detain suspects, or to collect or preserve evidence.[16] Doing so is, however, permitted only when other alternatives would not work. This provision reflects the reasoning of a constitutional case of 2018. In this case, the Constitutional Court of Korea held that a prosecutor's collection of the identities of mobile subscribers that accessed a single base station infringed the constitutional right to self-control of personal data and the freedom of communications and that doing so is thus unconstitutional.[17]

However, the previous MERS outbreak had shown the need for putting in place an effective contact tracing scheme when needed. This prompted an amendment of the Contagious Disease Prevention and Control Act (CDPCA)[18] so as to override the consent requirements under Korean data protection law in the event of an outbreak. There already is a provision in the PIPA, which exempts the application of the consent and other statutory requirement for temporary processing of personal data when there is an emergency need for public safety and security including public health.[19] The amendment of the CDPCA gave more concrete legal authority for implementing a contact tracing scheme during an outbreak of a contagious disease. After the onset of COVID-19, the Korean legislature further amended the CDPCA several times in order to better cope with the situations that had not been anticipated prior to the outbreak of COVID-19.

2. *Legal Basis for Centralized Contact Tracing*

For manual contact tracing by epidemiological investigators, interviews play a crucial role. Conducting interviews obviously takes time and sometimes accuracy could become an issue. As such, manual contact tracing has limitations in terms of the timely detection and quarantine of those suspected of being infected. Efforts were made in many parts of the world in order to make up for these limitations and several automated contact tracing models have been devised. Most of the newly devised models rely on geolocation data, typically gathered through smart phones. Each of these models has its own advantages and disadvantages as discussed below.

Depending on the provenance of the relevant data, these models can be divided into centralized models and decentralized models. There can also be a hybrid model. Among different types of automated contact tracing models, a majority of developed countries appear

[14] CSPA, Articles 18 and 19.
[15] As Korea has not signed the Budapest Convention on Cybercrime, there are several differences between the CSPA and wiretapping regimes of the US and EU.
[16] CSPA, Article 13(2).
[17] Constitutional Court of Korea, Case Ref. 2012 *Heonma* 538 (28 June 2018).
[18] Contagious Disease Prevention and Control Act [*Gamyeombyeongeu Yebang Mit Gwanri'e Gwanhan Beopryul*], Act No 17893 (last amended on 12 January 2021, effective as of 13 January, 2022).
[19] PIPA, Article 58(1)(3).

to have chosen decentralized 'privacy-preserving' proximity tracing models. These typically relay geolocation data utilizing the Bluetooth Low Energy technology. By design, these models grant data subjects the right to avoid tracking by not downloading or activating mobile apps. Soon after early efforts were made in order to develop and deploy a contact tracing model in the EU, the European Data Protection Board (EDPB) issued guidelines dated 21 April 2020. According to the EDPB guidelines, COVID-19 tracing apps would have to be based on the use of proximity data instead of geolocation data.[20]

For the decentralized approach, there are two subtypes: a fully decentralized approach and a partially decentralized approach. A fully decentralized approach works as follows. Through the operation of a mobile app, (i) smart phones exchange ephemeral IDs of individuals nearby via Bluetooth Low Energy ('Bluetooth Handshakes'); (ii) those individuals who are subsequently confirmed positive send their ephemeral IDs to a database in the server; and (iii) each app continues to download the database from the server and alerts if its owner has been in close proximity to one of those who are tested positive.[21] Apple-Google's Exposure Notification (AGEN) scheme is a well-known case of the decentralized approach.[22] AGEN has reportedly been embedded in the majority of European COVID-19 apps, including Austria's Stopp Corona, Germany's Corona-Warn-App, Italy's Immuni, Estonia's HOIA, the UK's NHS COVID-19, Protect Scotland, and StopCOVID NI (for Northern Ireland).[23] Japan also adopted AGEN in its contact tracing scheme called COCOA.

On the other hand, a main differentiating feature of the partially decentralized approach is that, in addition to being equipped with the functions of the fully decentralized app, a partially decentralized app would send ephemeral IDs collected from other smart phones to the server database so that it becomes possible to conduct contact tracing, risk analysis, and message transmission, utilizing the data accumulated at the server database.[24] Its examples include the Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) (specifically, the ROBERT protocol) and BlueTrace.[25] The PEPP-PT scheme was embedded in France's StopCovid and TousAntiCovid, and the BlueTrace approach was embedded in Singapore's TraceTogether and Australia's COVIDSafe.

Unlike these approaches, Korea has taken a centralized network-based contact tracing approach, which utilizes geolocation data collected from mobile carriers and other types of data

[20] EDPB, 'Guidelines 04/2020 on the Use of Location Data and Contact Tracing Tools in the Context of the COVID-19 Outbreak' (*EDPB*, 21 April 2020) https://edpb.europa.eu/our-work-tools/our-documents/ohjeet/guidelines-042020-use-location-data-and-contact-tracing-tools_en ('In the context of a contact tracing application, careful consideration should be given to the principle of data minimisation and data protection by design and by default: contact tracing apps do not require tracking the location of individual users. Instead, proximity data should be used; as contact tracing applications can function without direct identification of individuals, appropriate measures should be put in place to prevent re-identification; the collected information should reside on the terminal equipment of the user and only the relevant information should be collected when absolutely necessary.') Based on these Guidelines, the Norwegian Data Protection Authority (Datatilsynet), in June 2020, banned a GPS tracking COVID-19 app (named Smittestopp) which the Norwegian Institute of Public Health developed and released. Datatilsynet, 'Vedtak om midlertidig forbud mot å behandle personopplysninger – appen Smittestopp" (*Datatilsynet*, 6 July 2020) www.datatilsynet.no/regelverk-og-verktoy/lover-og-regler/avgjorelser-fra-datatilsynet/2020/vedtar-midlertidig-forbud-mot-smittestopp/.

[21] N Ahmed and others, 'A Survey of Covid-19 Contact Tracing Apps' (2020) 8 *IEEE Access* 134577 (hereafter Ahmed and others, 'A Survey of Covid-19').

[22] Apple and Google, 'Privacy Preserving Contact Tracing' (*Apple*, 2020). https://covid19.apple.com/contacttracing.

[23] PH O'Neill, T Ryan-Mosley, and B Johnson, 'A Flood of Coronavirus Apps are Tracking Us. Now It's Time to Keep Track of Them' (*MIT Tech Rev*, 7 May 2020) www.technologyreview.com/2020/05/07/1000961/launching-mittr-covid-tracing-tracker/.

[24] N Ahmed and others, 'A Survey of Covid-19' (n 21).

[25] Ibid.

that facilitate tracking of individuals. This approach does not allow its citizens to opt out of the contact tracing scheme. Only a few other jurisdictions, including Israel[26,27] and China,[28] appear to have taken this approach. In Korea, government agencies are granted a broad authority to process personal data during a pandemic for epidemiological purposes. Under the current provisions of the CDPCA, the Korea Disease Control and Prevention Agency (KDCA)[29] and municipal or/and local governments can, at the outbreak of an infectious disease, collect, profile, and share several categories of data that pertain to individuals who test positive or individuals who are suspected of being infected.[30] The data that can be collected include geolocation data; personal identification information; medical and prescription records (including the Drug Utilization Review [DUR]); immigration records; card transaction data for credit, debit, and prepaid cards; transit pass records for public transportation; and closed-circuit television (CCTV) footage.[31] In this context, 'individuals who are suspected to be infected' mean those who have been in close proximity to confirmed individuals, those who entered the country from a high risk region, or those who have been exposed to pathogens and other risk elements.[32] These individuals can be required to quarantine.[33] The CDPCA explicitly stipulates that the request of geolocation data under this law overrides the otherwise-applicable consent requirements under the LIA and CSPA.[34]

The KDCA can share the foregoing data with (i) central, municipal, or local governments, (ii) national health insurance agencies, and (iii) healthcare professionals and their associations.[35] The KDCA must also transfer a part of the data, including immigration records, card transaction data, transit pass records, and CCTV footage, to national health insurance information systems and other designated systems.[36]

Despite this legal mandate and authority, however, in practice, the scope and breadth of the data processed for contact tracing purposes and the recipients of the shared data have been much narrower, as explained in Subsections 3 and 4.

[26] Israel reportedly resorted to its emergency powers to redirect the counterterrorism monitoring program of the Israel Security Service (Shin Bet) into conducting contact tracing, which its Supreme Court later held to be unlawful unless the practice is permitted through legislation (Israeli Supreme Court, HCJ 2109/20, HCJ/2135/20, HCJ 2141/20 *Ben Meir v Prime Minister* (2020) (English translation) (*VERSA*, 26 April 2020) https://versa.cardozo.yu.edu/opinions/ben-meir-v-prime-minister-0).

[27] In July 2020, Israel's legislation, Knesset, passed a law authorizing the Security Service to continue to engage in contact tracing until 20 January 2021, and approved an extension of this period in January 2021(Knesset News, 'Foreign Affairs and Defense Committee approves continued use of the Shin Bet in the efforts to contain the spread of the coronavirus' (*The Knesset*, 13 January 2021), https://main.knesset.gov.il/EN/News/PressReleases/Pages/press13121q.aspx).

[28] China is also understood to have adopted a centralized approach utilizing QR codes, mobile apps, and other means, but its technical details have not been disclosed clearly (Paul Mozur et al., 'In Coronavirus Fight, China Gives Citizens a Color Code, with Red Flags' (*New York Times*, 7 August 2020), www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html).

[29] On 12 September 2020, the Korea Centers for Disease Control and Prevention (KCDC) was reorganized as a formal government agency to better combat the pandemic under the name of the Korea Disease Control and Prevention Agency (KDCA). References to the KDCA in this chapter include the agency's activities prior to the reorganization.

[30] CDPCA, Article 76-2.

[31] CDPCA, Article 76-2(1)(2).

[32] CDPCA, Article 2(xv-2).

[33] CDPCA, Article 42(1).

[34] CDPCA, Article 76-2(2).

[35] CDPCA, Article 76-2(3).

[36] CDPCA, Article 76-2(4).

3. Legal Basis for QR Code Tracking

The amendment to the CDPCA of 4 March 2020[37] authorized the KDCA, the Ministry of Health and Welfare, and municipal and/or local governments to issue decrees to citizens 'to keep the list of administrators, managers, and visitors at the venues or facilities having the risk of spreading infectious diseases'.[38] This new provision enabled the KDCA to deploy an electronic visitor list system by utilizing QR (quick response) codes.

4. Legal Basis for the Disclosure of the Routes of Confirmed Cases

Under the CDPCA, at the outbreak of a serious infectious disease, the KDCA and municipal and/or local governments must promptly make the following information publicly available on the Internet or through a press release: the path and means of transportation of confirmed cases; the medical institutions that treated the cases; and the status of relevant close contacts.[39] Anybody can appeal if the disclosed information is incorrect or if there is any opinion. From this appeal, if deemed needed, the KDCA or municipal and/or local governments should immediately take necessary remedial measures such as making a correction.[40]

This provision allowing for public disclosure of information is an important exception to the principles set forth under Korea's data protection laws. This provision was introduced in the CDPCA in 2015 following the MERS outbreak. At the time, epidemiologists first requested the government to disclose the information about the hospitals that treated confirmed cases and also about the close contacts in order to protect healthcare professionals from the risk of infection.[41] The public opinion also urged the government to ensure transparency by disclosing whereabouts of confirmed cases.[42] In response, the government disclosed the list of the hospitals that treated confirmed cases on 5 June 2015, breaking the non-disclosure principle for the first time. A bill for the foregoing provision was submitted on the same day and was passed by the legislature on 6 July 2015.[43] The bill was passed within a very short period of time and, as such, there was insufficient time to consider and debate privacy concerns and other important implications that would arise from the amendment. Following the outbreak of COVID-19 in 2020, this provision was immediately triggered, raising considerable privacy concerns as explained below in Sub-section V 2.

5. Legal Basis for Quarantine Monitoring

The amendment to the CDPCA of 4 March 2020[44] authorized the KDCA and municipal and/or local governments to check the citizens for symptoms of infectious diseases and to collect geolocation data through wired or mobile communication devices.[45] This new provision enabled the KDCA to track GPS data to monitor those quarantined at home.

[37] Effective as of 5 June 2020.
[38] CDPCA, Article 49(1)(ii-ii).
[39] CDPCA, Article 34-2(1).
[40] CDPCA, Article 34-2(3)(4).
[41] The Korean Society of Infectious Diseases, 'White Paper on Chronicles of MERS' (*KSID*, 2015) www.ksid.or.kr/file/mers_170607.pdf.
[42] Ibid.
[43] Effective as of 7 January 2016.
[44] Effective as of 5 June 2020.
[45] CDPCA, Article 42(2)(ii).

Prior to this amendment, the quarantine monitoring app had already been in use. During this period, in order to comply with the consent requirements for the collection and use of personal geolocation data under the LIA, the app to be used for monitoring purposes made a request to an installer to click on the consent button before installation process starts. Because installing the monitoring app and providing the requisite consent allowed one to avoid the inconvenience of being manually monitored by the quarantine authorities or of facing the possibility of being denied entry into the country, most individuals who were subject to quarantine appear to have chosen to use the app. It was not entirely clear whether such involuntary agreement to download and activate the app constitutes valid consent under the LIA, and the foregoing amendment to the CDPCA clarified the ambiguity by explicitly allowing the collection of geolocation data for quarantine monitoring purposes.

III. ROLE OF TECHNOLOGY IN KOREA'S RESPONSE TO COVID-19

A variety of technological means were employed in the process of coping with the pandemic in Korea. Among these, the most important means would include the tools to gather and utilize geolocation data for the purposes of engaging in contact tracing and other tracking activities. The following describes how technological tools were deployed.

1. *Use of Smart City Technology for Contact Tracing*

Based on the mandate and authority under the CDPCA, the Korean government launched the COVID-19 Epidemic Investigation Support System (EISS) on 26 March 2020.[46] By swiftly remodeling the EISS from the existing smart city data hub system developed by several municipal governments, Korea could save time during early days of the pandemic. Prior to the outbreak of COVID-19, in accordance with the Smart City Act,[47] the Korean central and municipal and/or local governments had been developing and implementing smart city hubs; several 'smart cities' have been designated as test beds for innovation in an effort to foster the research and development in areas related to sharing-economy platforms, AI services, Internet-of-Things technologies, renewable energy, and other innovative businesses. In relative terms, compared to a situation in which systems developed for security service agencies are redeveloped and used for contact tracing purposes, the use of a smart city system might have the advantage of heightened transparency and auditability.

The EISS collects requisite data pertaining to confirmed cases and those who are suspected to have been in contact. Data that can be collected includes base station data from mobile carriers and credit card transaction data from credit card companies. In order to obtain data, clearances should be obtained from the police and from the Credit Finance Association (CREFIA), respectively, for base station data and for credit card transaction data. After clearances are obtained, transfer of the data to epidemiological investigators takes place on a near real-time basis.[48] Equipped with base station data and credit card transaction data, epidemiological investigators can effectively track many of the confirmed cases and their close contacts, as

[46] A pilot operation started on 16 March 2020.

[47] The Act on Construction of Smart Cities and Industry Promotion [*Smart Doshi Joseong Mit San'eop Jinheung Deung'e Gwanhan Beopryul*], Act No 17799 (last amended on 29 December 2020, to be effective as of 30 December 2021).

[48] The Ministry of Land, Infrastructure and Transport (MOLIT), 'Online Q&A for the Support System for the COVID-19 Epidemiological Investigation' (*MOLIT*, 10 April 2020), www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?id=95083773.

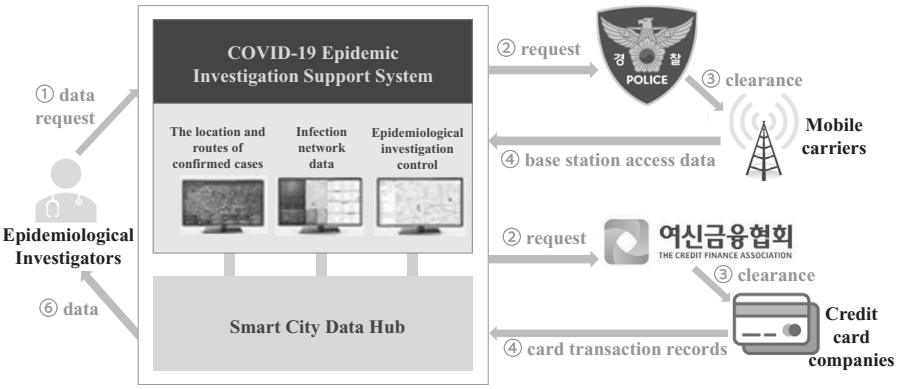

FIGURE 18.2 The COVID-19 Epidemic Investigation Support System
Note: MOLIT, 'COVID-19 Smart Management System (SMS), formally named 'COVID-19 Epidemic Investigation Support System (EISS)' (*MOLIT*, 6 December 2020), https://smartcity.go.kr/ (hereafter MOLIT, 'COVID-19 Smart Management System').

Korea is reported to have the highest penetration rate in the world for mobile phones and for smart phones, respectively at 100% and 95% as of 2019 (Figure 18.2).[49]

In addition to the EISS, epidemiological investigators at municipal or local governments can, upon request, be given access to the DUR by the KDCA. Under 'normal' circumstances, a main use of the DUR would be to give useful information about various drugs to the general public and to those engaged in the pharmaceutical supply chain. In the context of COVID-19, the DUR could further be used for obtaining requisite tracing data.

2. *Use of QR Codes for Tracking Visitors to High-Risk Premises*

On 10 June 2020, shortly after the 2020 amendment to the CDPCA came into force, Korea further launched a QR code-based electronic visitors' log system to track visitors to certain designated types of high-risk premises such as restaurants, fitness centers, karaoke bars, and nightclubs. This system was deployed with the help of two large Internet platform companies, Naver and Kakao, and of mobile carriers through an app called Pass (Figure 18.3).

With this system in place, for instance, a visitor to a restaurant must get an ephemeral QR code pattern from a website or mobile app provided by the Internet platform companies or mobile carriers, and have the pattern scanned using an infrared dongle device maintained by the restaurant, typically at the entrance.[50] That way, QR code-based electronic visitor lists are generated and maintained for these premises (KI-Pass). Maintaining this tracking system could, however, raise concerns over privacy or surveillance. In order to address these concerns, identifying information about the visitors is kept separately from the information about individual business premises. More details about this bifurcated system are provided in Sub-section IV 2.

[49] Pew Research Center, 'Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally' (*Pew research*, 5 February 2019), www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/.

[50] MOHW, 'Guidance on the Use of Electronic Entry Lists (for Visitors and Managers)' (*NCOV*, 10 June 2020), http://ncov.mohw.go.kr/shBoardView.do?brdId=2&brdGubun=25&ncvContSeq=2603 (hereafter MOHW, Guidance on the Use of Electronic Entry Lists).

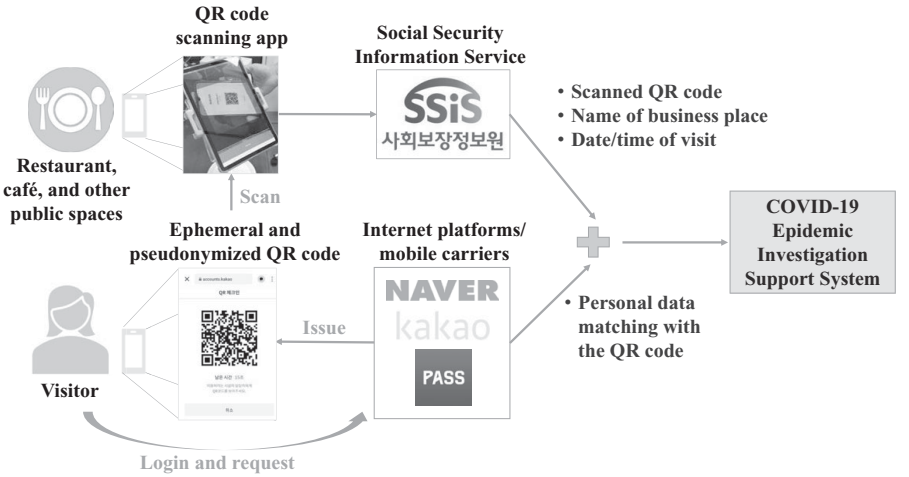

FIGURE 18.3 The KI-Pass, a QR code-based electronic visitor booking system
Note: Naver Corporation, 'QR Check-In' (*NAVER*, 2020) https://m.help.naver.com/support/contents.

3. *Public Disclosure of the Routes of Confirmed Cases*

Routes of confirmed cases are disclosed on the websites of the relevant municipal and/or local governments, in a text or tabular form. No enhanced technology is used for the disclosure. The disclosed information is also sent to mobile phones held by nearby residents as an emergency alert message in order to alert them of the possible exposure and risks.

4. *Use of GPS Tracking Technology and Geographic Information System (GIS) for Quarantine Monitoring*

The CDPCA also grants authorization for quarantine measures to government agencies. Thus, a 14-day quarantine requirement was introduced for (1) individuals who are deemed to have been in close proximity to confirmed cases[51] and (2) individuals who arrive from certain high-risk foreign countries.[52] To monitor compliance, those who are under quarantine are required to install and run a mobile app called the 'Self-Quarantine Safety Protection App' developed by the Ministry of the Interior and Safety. The app enables officials at competent local governments to track GPS data from smart devices held by those quarantined on a real-time basis, through the GIS, in order to check and confirm whether they have remained in their places of quarantine. Also, quarantined individuals are expected to use the app to report symptoms, if any, twice a day (Figure 18.4).

IV. FLOW OF DATA

In a nutshell, developing and deploying a tracing system is about gathering and analyzing data. While using the collected data for epidemiological purposes could be justified on the basis of public policy reasons, legitimate concerns over surveillance and privacy could be raised at the

[51] Implemented from 23 February 2020.
[52] Expanded to all countries as of 1 April 2020.

FIGURE 18.4 User interface of the Self-Quarantine App
Note: Google Play Store and Ministry of the Interior and Safety, Self-Quarantine Safety Protection App, https://play.google.com/store/apps/

same time. As such, it is imperative to consider provenance and governance of various types of data. A starting point for doing this would be to analyze the flow of data, to which we now turn.

1. *Centralized Contact Tracing*

Personal data including geolocation data of an individual flows within the EISS in the following steps: (i) the KDCA or municipal and/or local governments make a request; (ii) the police and/or the CREFIA give clearances to the transfer of mobile base station data and/or credit card transaction data, respectively; (iii) mobile carriers and/or credit card companies provide data as requested; (iv) epidemiological investigators review and analyze data pertaining to confirmed cases; (v) the investigators verify and obtain further information through interviews with confirmed cases; (vi) the investigators further conduct epidemiological network analysis and identify epidemiological links regarding the spread of COVID-19; and (vii) the KDCA and municipal and/or local governments receive relevant data and implement necessary measures such as quarantine or the disinfection or shutdown of premises where confirmed individuals visited.[53]

[53] MOLIT, 'COVID-19 Smart Management System (SMS) <Formally Named 'Epidemic Investigation Support System (EISS)'>' (*MOLIT*, 6 December 2020) https://smartcity.go.kr/2020/06/12/%ec%bd%94%eb%a1%9c%eb%82%9819-%ec%97%ad%ed%95%99%ec%a1%b0%ec%82%ac-%ec%a7%80%ec%9b%90%ec%8b%9c%ec%8a%a4%ed%85%9c-%ec%84%a4%eb%aa%85%ec%9e%90%eb%a3%8c-%eb%b0%8f-qa/ (hereafter MOLIT, 'COVID-19 Smart Management System').

In the whole process, mobile base station data plays a crucial role for tracing purposes. Mobile base station data contains the names and phone numbers of the individuals who were near a specific base station. Exact location data were not collected, although collecting such data would have been technically feasible through triangulation using latitude or longitude data. However, as mobile base stations are installed at an interval of 50 to 100 meters in a downtown of a densely populated city such as Seoul, base station data can be considered precise enough for the purpose of identifying those who stayed near a confirmed case. At the same time, because the geographic coverage of a base station could be rather broad, there could be an issue of over-inclusion, with implications on privacy.

2. *QR Code Tracking*

An outbreak in May 2020 was investigated and found to have an epidemiological relationship to a night club located in the Itaewon district, in Seoul. When this outbreak became serious, efforts were made to locate the individuals affected and to conduct interviews so that further preventative measures could be deployed. However, only 41.0% of individuals, (i.e., 2,032 out of 4,961 individuals), could be contacted by epidemiological investigators over the phone.[54] This was mainly due to the fact that the visitor list was hand-written by the visitors themselves and that sexual minorities who visited the club wrote down false identifies and/or phone numbers for fear of being forced to reveal their sexual orientations. This inability to contact a larger number of visitors to a particular premise reinforced the view that paper visitor lists should be substituted, where possible, with electronic visitor lists, so that the accuracy of the information contained in the visitor lists can be all but guaranteed.

This hastened the development of a QR code-based electronic visitor list system, which was deployed on 10 June 2020.[55] When a visitor has his or her QR code scanned by an infrared dongle device installed at a business premise, the manager of the business premise does not collect any personal data, other than the code itself. Under the system deployed in Korea, visitor identification information is held by the issuers of QR codes only, unless a need arises to confirm the identity for epidemiological purposes. Specifically, one of the three private entities which issue QR codes holds visitor identification information: Internet platform companies Kakao and Naver and mobile carriers who jointly developed the app named Pass. Data directly related to business premises are held by the Social Security Information Service (SSIS). That is, the SSIS collects the following data: the name of the business premise, time of entry, and encrypted QR codes. The SSIS does not hold any personally identifiable data in this context.[56] That way, relevant data are kept separately, and a bifurcated system is maintained. When a report is made that a visitor to a business premise is confirmed positive, the bifurcated datasets are then combined on a need basis in order to retrieve the relevant contact information, which is transmitted to the EISS. The transmitted information is then used by the KDCA and municipal and/or local governments for epidemiological investigations. The data generated by QR-code scanning is automatically erased after four weeks.[57]

[54] MOHW, Guidance on the Use of Electronic Entry Lists (n 50).
[55] Ibid.
[56] Ibid.
[57] Ibid.

TABLE 18.1. *11 January 2021 Disclosure of the local government of Gwanak-gu, Seoul*[58] *(case numbers redacted)*

□ Status of Case No. ****
- Source of Infection: Presumably infected from a family member
- Confirmed positive on 11 January.
□ Status of Case No. ****
- Source of Infection: Presumably infected from a family member
- Confirmed positive on 11 January.
□ Status of Case No. ****
- Source of Infection: Presumably infected from a confirmed case at the same company in a different region
- Confirmed positive on 11 January.
□ Status of Case No. ****
- Source of Infection: Under investigation
- Confirmed positive on 11 January.
□ Status of Case No. ****
- Source of Infection: Under investigation
- Confirmed positive on 11 January.
□ Status of Case No. ****
- Source of Infection: Under investigation
- Confirmed positive on 11 January.
※ Measures
- Will transfer confirmed cases to the government-designated hospitals
- Will disinfect the residence and neighboring areas of confirmed cases
- Investigating visited places and close contacts

3. *Public Disclosure of the Routes of Confirmed Cases*

As explained above, municipal and/or local governments receive geolocation data and card transaction data from the EISS and disclose a part of the data to the general public. At an earlier stage of the COVID-19 outbreak, very detailed routes of confirmed cases were disclosed to the public. These disclosures did not include the names or other personally identifiable information of the confirmed individuals. What was revealed typically included a pseudonym or part of the full name of the infected individual as well as sex and age. In addition, vocation and/or area of residence was often disclosed. Although directly identifiable personal information was not disclosed, sometimes simple investigation and profiling would enable re-identification or reveal personal details. Certain individuals indeed became subject to public ridicule, after their identities were revealed. Debates on privacy followed, and the KDCA revised its guidelines about public disclosure of information on contact tracing. As a result, municipal and/or local governments are now disclosing much more concise information focusing on locations and premises rather than on an individuals' itinerary. Also, disclosure information is deleted after fourteen days following disclosure. One of the examples is as shown in *Table 18.1*.

4. *Quarantine Monitoring*

The self-quarantine app collects GPS data from mobile devices and shares it with the GIS, so that an official at the local government can monitor the location of a quarantined individual on a real-time basis.

[58] Gwanak-gu Local Government, 'The Statuses and Routes of COVID-19 Confirmed Cases' (*Gwanak*, 11 January 2021) www.gwanak.go.kr/site/health/ex/bbs/View.do?cbIdx=587&bcIdx=117494&parentSeq=117494.

V. DATA FLOW AND DATA GOVERNANCE

Collecting relevant data on a near real-time basis is crucial in order to contain the spread of COVID-19. At the same time, data collection immediately raises privacy concerns. As such, a delicate balance must be struck between conducting effective epidemiological investigations and protecting the privacy of individuals. Delineating the precise flow and provenance of collected data will give implications as to how the delicate balance can be struck and maintained. The following section gives some explanations as to what transpired in Korea in this respect.

1. *Centralized Contact Tracing (including QR Code Tracking)*

An early response is critical to contain the spread of highly infectious diseases such as COVID-19. In turn, the effectiveness of such a response relies on the prompt collection and sharing of accurate data about confirmed cases and close contacts. Manual epidemiological tracing has serious limitations. It takes time for human investigators to conduct manual tracing, causing delays. Also, manual tracing is vulnerable to faulty memory or deception on the part of interviewees, resulting in inaccurate epidemiological reports.

In response to the rapid spread of COVID-19, Korea chose to integrate such human efforts with a technology-driven system of data processing. For example, a prompt compilation of geolocation data has been a crucial enabling factor in Korea's contact tracing strategy. The EISS, which makes use of the smart city technology, allowed public health authorities to efficiently allocate valuable resources. With the assistance of technology, for instance, epidemiological investigators were able to conduct tracing in a more effective and efficient manner.

At the same time, questions were raised whether the centralized contact tracing model adopted in Korea was overly intrusive, even harmful to fundamental freedoms constituting the very cornerstones of a democratic society. The collection of data has sometimes been equated to mass surveillance, raising privacy concerns as well. This line of criticism would have a clear merit, if certain other alternative tracing systems show the same or even higher level of efficacy, while collecting less granular and less detailed personal data.

The problem, however, is that, while a decentralized system such as the Bluetooth-based approach is in general better in protecting privacy, it has its own shortcomings that are yet to be solved. First, a tracing app needs to attain a certain penetration rate, in other words, the proportion of active users of the mobile app among the whole population should be sufficiently high for a tracing system to function properly. In order to achieve the so-called digital herd immunity this penetration rate should be fairly high – sometimes set at 60 to 75%.[59] To date, most countries have failed to achieve this level of penetration rate, due to, among other things, low levels of smartphone penetration rates. Second, Bluetooth-based proximity tracing may not work effectively in crowded areas that are in fact prone to experience explosive outbreaks of infectious diseases such as COVID-19. Third, decentralized models generally do not allow for human-in-the-loop based verification and tend to show excessively high false positives.[60] Fourth, iOS does not allow third-party apps running in the background to function properly in order to

[59] V B Bulchandani and others, 'Digital Herd Immunity and COVID-19' (2020) https://arxiv.org/pdf/2004.07237.pdf.

[60] J Bay and others, 'BlueTrace: A Privacy-Preserving Protocol for Community-Driven Contact Tracing across Borders' (*Government Technology Agency*, 9 April 2020) https://bluetrace.io/static/bluetrace_whitepaper-938063656596c104632def383eb33b3c.pdf.

broadcast Bluetooth signals, unless the AGEN system is deployed.[61] Fifth, for a fully decentralized approach, there would be no informational benefits to public health authorities because relevant information simply does not flow to public health authorities. While this could be beneficial in maintaining the privacy interests of citizens, at the same time, precious opportunities for gaining epidemiological data would be lost. Lastly and perhaps most fundamentally, the decentralized approach has to rely on good-faith cooperation by confirmed individuals. That is, the approach would not work unless confirmed individuals make voluntary reports and, as such, this approach exhibits a similar problem as in a manual tracing method.

This is not to say that a centralized model would always be preferable. While a decentralized approach may not lead to a herd immunity, it could nonetheless play a complementary role in containing the spread of COVID-19, particularly in densely populated areas such as city centers and on university campuses. Thus, as a general matter, a centralized approach and a decentralized approach each have their own strengths and limitations. For a centralized approach, its main strengths would include: immediate availability undeterred by the penetration levels; effective response to mass infection; no compatibility concerns; and most importantly, impactful contribution to epidemiological investigations.

In the case of Korea, there is no denying that contact tracing and other tracking mechanisms were a crucial component in the whole apparatus dealing with daunting challenges caused by COVID-19. Also, overall, Korean society at large complied with the requirements imposed by these tracking mechanisms, without raising serious privacy concerns. If we go one step further, there could be various viewpoints and reactions as to why Korean citizens in general complied with the measures adopted by the government.

In terms of data protection, the PIPA was enacted in 2011 and earlier statutes also contained various elements of data protection. Separately, the Constitutional Court of Korea, in 2005, declared that the right to data protection is a constitutional right. As such, Korean citizens are in general well aware of the value of data protection in modern society. In adopting technology-based contact tracing mechanisms and complying with the requirements associated with such contact tracing mechanisms, it can be said that the Korean society as a whole made a wide-ranging value judgement about privacy, public health, and other social and legal values. Among other things, citizens exhibited a striking willingness to cooperate with authorities in their efforts to collect epidemiological data including geolocation data, which can be traced back to their previous experience with the MERS outbreak. Utilizing new technologies for epidemiological purposes was perhaps not much of an additional concern as, in relative terms, many Koreans are at ease with adapting to new technological environments.

This does not mean, however, that the data collection was without controversy. On the contrary, several activist groups joined forces and filed a constitutional petition seeking the Constitutional Court of Korea's decision regarding the constitutionality of contact tracing mechanisms.[62] More specifically, the petition challenges the constitutionality of the CDPCA provisions which enabled contact tracing in the first place.[63] It also views the government's collection of mobile base station data based on these provisions unconstitutional, in particular pointing to the collection of data about the visitors at a night club in the Itaewon area during an outbreak, as doing so violates, among other things, the constitutional right to self-determination of personal

[61] J Taylor, 'Covidsafe App Is Not Working Properly on iPhones, Authorities Admit' (*The Guardian*, 6 May 2020), www.theguardian.com/world/2020/may/06/covidsafe-app-is-not-working-properly-on-iphones-authorities-admit.

[62] Constitutional Court of Korea, Case Ref. 2020 *Heonma* 1028 (filed on 29 July 2020, pending).

[63] CDPCA, Articles 2-15, 76-2.

data.[64] This petition grounds itself on the Court's 2018 decision that held the collection of the identities of mobile subscribers that accessed a particular base station in the course of criminal investigation unconstitutional.[65] Regardless of the outcome of this case, the scope of geolocation data might need to be adjusted to balance epidemiological benefits with privacy.

In the case of QR codes, the bifurcated approach perhaps helps mitigate security risks and privacy concerns, by separating personally identifiable data from visitor logs and by combining them only when necessary for epidemiological investigation. Also, while both a paper form for visitor logs and a QR-code based electronic log system are usually available at business premises, the general public appears to prefer the QR code-based electronic visitor log system. Part of the reason would be the trustworthiness of the QR-code system. That is, while there is virtually no concern over possibilities of data breach for a QR code-based electronic visitor log system, a paper visitor list could be vulnerable to illegal leakage by employees of business premises or by subsequent visitors.[66]

About the contact tracing mechanism in general, there could be a concern over possibilities of 'function creep.' The concern is that, while conducting contact tracing under the current extraordinary circumstances of COVID-19 could be justified, after the pandemic is over, the government may be tempted to use this mechanism for surveillance purposes. In the case of Korea, there are two built-in safeguards against this from happening. First, data collection for epidemiological purposes is under the sole purview of the KDCA and the relevant databases are maintained by the KDCA as well. This means that, even if the government is tempted to divert the system for different purposes, doing so would be a cumbersome procedure simply because the system is maintained and held by a single public health agency with a narrow public health mandate. Second, the KDCA's authority for the current data collection is, for the most part, derived from statutory provisions contained in the CDPCA and not from the PIPA, a general data protection statute. After the pandemic is over, the KDCA or any other government agencies would require a separate statutory rationale in order to collect data.

Compared to the Korean government's active role in utilizing technology to cope with the COVID-19 pandemic, public-private collaboration based on the sharing of public data and the use of open APIs (application programming interfaces) in Korea has somewhat lagged. There have been recent cases of meaningful contributions from the private sector, however. An example would be a collaborative dataset sourced from public disclosures, which has been actively used for visualization purposes and also for machine learning training purposes.[67]

2. *Public Disclosure of the Routes of Confirmed Cases*

Unlike contact tracing itself, which was generally accepted as a necessary trade-off between privacy and public health in facing the pandemic, the public disclosure of the routes of

[64] Joint Representatives for the Petition for the Decision that Holds COVID-19 Mobile Base Station Data Processing Unconstitutional, 'Petition' (*Opennet*, 29 July 2020) https://opennet.or.kr/18515.

[65] Constitutional Court of Korea, Case Ref. 2012 *Heonma* 538 (28 June 2018).

[66] MOHW, 'Guidance on the Use of Electronic Entry Lists' (n 50).

[67] J Kim and others, 'Data Science for COVID-19 (DS4C)' (*Kaggle*, 2020), www.kaggle.com/kimjihoo/coronavirusdataset/data. Another example is SK Telecom's support of an AI-based teleconference system for quarantine monitoring: ZDNET, 'SKT Reducing COVID-19 Monitoring Workloads up to 85% Using AI' (*ZDNET*, 25 June 2020), https://zdnet.co.kr/view/?no=20200625092228. Refer to CHOSUNBIZ, 'Taking up to 30,000 Calls a Day When 2,000 was a Challenge Due to the Coronavirus … "Thank you AI"' (*ChosunBiz*, 24 May 2020), https://biz.chosun.com/site/data/html_dir/2020/05/23/2020052301886.html for other Korean examples of private initiatives utilizing AI related to the COVID-19 pandemic.

confirmed cases quickly became controversial due to privacy concerns. Such public disclosures were, in fact, another policy response from the experiences of the MERS outbreak. That is, during the MERS outbreak, there was great demand for transparency and some argued that the lack of transparency impeded an effective response. However, with the onset of the COVID-19 outbreak, the pendulum swung in the other direction. Not just the detailed nature but also the uneven scope and granularity of disclosures among the KCDA and the numerous municipal and local authorities caused confusion, in particular during the initial phase. Concerns were not limited to the invasion of privacy. Private businesses, such as restaurants and shops, that were identified as part of the routes often experienced abrupt loss of business.

These concerns were encapsulated in the recommendation issued by the National Human Rights Commission (NHRC) on 9 March 2020.[68] The NHRC expressed concerns about unwanted and excessive privacy invasion as well as secondary damages such as public disdain or stigma, citing a recent survey showing that the public was even more fearful of the privacy invasion and stigma stemming from an infection than the associated health risk itself.[69] The NHRC noted that excessive public disclosure could also undermine public health efforts by dissuading those suspected of infection from voluntarily reporting their circumstances and/or getting tested for fear of privacy intrusions.[70] The NHRC further recommended that route disclosures be made in an aggregate manner focusing on locales at issue, rather than disclosing the times and places of visits at an individual level and possibly revealing personal itineraries.[71]

In response to the NHRC's recommendations, the KDCA issued its first guidelines regarding public disclosures to municipal and local governments on 14 March 2020, which limited the scope and detail of the information to be made publicly available. Specifically, the KDCA (i) limited the period of route disclosure from one day prior to the first occurrence of symptoms to the date of isolation, (ii) limited the scope of visited places and means of transportation to those spatially and temporally proximate enough to raise concerns of contagion, considering symptoms, duration of a visit, status of contacts, timing, and whether facial masks were worn, and (iii) banned the disclosure of home addresses and names of workplaces. On 12 April 2020, the KDCA further revised the guidelines. Under the revised guidelines, (i) information on routes should be taken down 14 days after the confirmed case's last contact with another individual, (ii) information on 'completion of disinfection' should be disclosed for relevant places along the disclosed routes, and (iii) the period of route disclosure should start from two days prior to the first occurrence of symptoms.[72] One complication from public disclosures of information is that, once a disclosure is made, the disclosed information is rapidly further disseminated via various social media outlets by individual users. Thus, data protection agencies have been actively sending out takedown notices to online service providers to ensure that such content is taken down following the 14-day period.

In May 2020, a spate of confirmed cases arose at a nightlife district in Itaewon, Seoul, that is frequented by persons with a specific sexual orientation. While public health authorities mounted a campaign urging prompt testing for those who could be at risk, it was ostensible

[68] NHRC, 'Statement Concerning the Excessive Disclosure of Private Information Pertaining to Confirmed COVID-19 Cases' (*NHRC*, 9 March 2020), www.humanrights.go.kr/site/program/board/basicboard/view?currentpage=2&menuid=001004002001&pagesize=10&boardtypeid=24&boardid=7605121.

[69] Ibid.

[70] Ibid.

[71] Ibid.

[72] KDCA, 'Guidance to Information Disclosure of Transit Routes of Confirmed Patients, etc.' (*KDCA*,12 April 2020), www.cdc.go.kr/board.es?mid=a20507020000&bid=0019&act=view&list_no=367087.

that the fear of being forced to reveal sexual orientations or being socially ostracized was a significant deterring factor. In response, the Seoul Metropolitan government initiated anonymous testing from 11 May 2020, under which individuals were only asked for their phone numbers. The anonymous testing scheme expanded and began to be applied to the whole country on 13 May 2020.

After witnessing these debates, the KDCA issued further revised guidelines dated 30 June 2020. The latest guidelines provided that municipal and/or local governments should disclose the area, the type of premises visited, the trade names and addresses of these premises, the date and time of exposure, and disinfection status and that disclosures should not be made for each individual and his or her timeline but instead in the format of 'lists of locations visited.' The guidelines further stipulated not to disclose information regarding the visited places if all close contacts have been identified.[73]

Subsequently, an amendment to the CDPCA was made dated 29 September 2020 and this amendment, among others, included a provision that excludes from the scope of public disclosure the 'sex, age, and other information unrelated to the prevention of contagious disease as stipulated in the Presidential Decree.'[74] The current Presidential Decree for the CDPCA lists the name and detailed address as examples of such 'other information unrelated to the prevention of contagious disease.'[75]

The above shows the ongoing process of trial and error in search of a more refined approach which would better balance the imperatives emanating from public health concerns during a pandemic with privacy and other social values. Urgency of the situation perhaps made it imperative to implement swift measures for gathering information. While implementing swift measures is inevitable, it is also important to review the legitimacy and efficacy of these measures on an ongoing basis and to revise if needed. For instance, compared to the disclosure of precise routes profiled for each confirmed case, the disclosure of aggregated route information has proven sufficient to achieve the intended public health policy goals. As demonstrated in the *Itaewon Case*, a less privacy-intrusive alternative can also assist infection control efforts by encouraging voluntary reporting and testing.

Regarding the disclosure of the names and addresses of business premises, assuming that disinfection can effectively address contagion risks, the only benefit would be to alert other visitors and to encourage them to self-report and get tested. Therefore, if all visitors are in fact identifiable through contact tracing, the public disclosure of the type of business and the broader area of the location, rather than identifying the name of the specific business premise, would be sufficient for purposes of public health. In fact, revisions to the KDCA guidelines were made reflecting practical lessons learned throughout 2020 and provide for deletion of data that is unnecessary or no longer necessary.

3. *Quarantine Monitoring*

Human surveillance of quarantined persons is often costly, ineffective, and in many cases inevitably intrusive. The quarantine monitoring through GPS tracking has generally been

[73] KDCA, 'Guidance to Information Disclosure of Transit Routes of Confirmed Patients, etc.' (3rd ed) (30 June 2020), www.gidcc.or.kr/wp-content/uploads/2020/02/%ED%99%95%EC%A7%84%EC%9E%90_%EB%8F%99%EC%84%A0_%EB%93%B1_%EC%A0%95%EB%B3%B4%EA%B3%B5%EA%B0%9C_%EC%95%88%EB%82%B43%ED%8C%90.hwp.

[74] CDPCA, Article 34-2 (1).

[75] Presidential Decree for CDPCA, Article 22-2 (1).

regarded as a more effective but less intrusive substitute for the human surveillance. As such, there have not been serious privacy concerns raised about quarantine monitoring.

4. *Data Governance*

On a regulatory front, the outbreak of COVID-19 has highlighted the need for Korea's privacy and data protection authorities to be ever more vigilant during public emergencies. In February 2020, Korea undertook a major reform to its privacy and data protection laws which came into effect as of 5 August 2020. As a result of the amendments, Korea's data protection authority will be consolidated and vested in the Personal Information Protection Commission (PIPC). This reform is expected to allow the PIPC to engage in a more proactive role in balancing the rights of data subjects with public health goals and to provide clearer guidance as to what to disclose and how to de-identify when making public disclosure.

On a broader level, in terms of the flow and provenance of data, two general directions can be distinguished. One direction is from the general population to public health authorities. Data gathered and shared in this direction is mainly done in order to carry out contact tracing, to conduct epidemiological analyses, and to devise and implement public health measures. At the same time, data flows toward the other direction as well, from the government and public health authorities to the general public. What is carried out in this context is mostly public disclosures of data about confirmed cases. Doing this would presumably be helpful for purposes of enhancing transparency and giving alerts so that citizens can prepare.

Regarding both directions of data flows, there are tensions between public health purposes and privacy interests: gathering and disseminating detailed information would in general be helpful in containing the spread of COVID-19, while, at the same time, doing so could be detrimental to the protection of the privacy of citizens. Details of the tensions, however, are different between the two directions of data flows. When data flows from the general public to public health authorities, a major concern would be the possibility of surveillance. Seen from a public policy perspective, attention would thus need to be paid as to whether and how a possible concern over surveillance could be assuaged. Putting in place systematic and procedural safeguards could be helpful. On the other hand, when data flows from public health authorities to the general public, mostly in the form of public disclosures of data about confirmed cases, concerns could be raised about the privacy of citizens. A privacy concern in this context could arise due to the possibility of the revelation of unwanted or embarrassing personal details. The risk could be elevated, if there is an added motivation for a public officer to gain attention through media, by leaking a 'headline grabbing' news item. In that regard, attention may need to be paid as to what data is made available to public sector officers.

VI. LOOKING AHEAD

As the COVID-19 outbreak continues its course, new societal challenges or existing ones that are being exacerbated by the pandemic such as the digital divide, are gathering more attention in Korea and elsewhere. Heightened concerns of ostracization or stigma directed to minority groups, the vulnerability of health and other essential workers that face constant exposure to infections, and children from underprivileged families that are ill equipped for remote learning are but a few examples. The *Itaewon case*, discussed earlier, has demonstrated the need for authorities to be prepared to promptly address concerns of prejudice against minority groups in the Korean society. The same should be said regarding the acute health and economic

disadvantages faced by the underprivileged during a pandemic. Yet, the societal challenges in the post-COVID-19 era, with its trend towards remote work, education, and economic activity will likely call for more long-term and fundamental solutions.

In this regard, the active use and application of AI and data analytics, as well as a robust ethical review concerning its governance, is expected to be critical in achieving the social reforms required to cope with the challenges of the present and coming future. In doing so, a pre-requisite would be to compile and draw a 'data map' so that data's flow and provenance can systematically be understood. With such understanding, further discussions could perhaps be made regarding appropriate levels of granularity for data disclosures and different levels of access control and other safeguards, depending on specific needs or policy goals. Korea's experience dealing with COVID-19 can provide a valuable lesson in this context.

PART VI

Responsible Corporate Governance of AI Systems

19

From Corporate Governance to Algorithm Governance

Artificial Intelligence as a Challenge for Corporations and Their Executives

Jan Lieder

I. INTRODUCTION

Every generation has its topic: The topic of our generation is digitalization. At present, we are all witnessing the so-called industrial revolution 4.0.[1] This revolution is characterized by the use of a whole range of new digital technologies that can be combined in a variety of ways. Keywords are self-learning algorithms, Artificial Intelligence (AI), autonomous systems, Big Data, biometrics, cloud computing, Internet of Things, mobile internet, robotics, and social media.[2]

The use of digital technologies challenges the law and those applying it. The range of questions and problems is tremendously broad.[3] Widely discussed examples are self-driving cars,[4] the use of digital technologies in corporate finance, credit financing and credit protection,[5] the

[1] For details, see Bundesministerium für Bildung und Forschung, 'Zukunftsbild Industrie 4.0' (*BMBF*, 30 December 2020) www.bmbf.de/bmbf/de/forschung/digitale-wirtschaft-und-gesellschaft/industrie-4-0/industrie-4-0.html; P Bräutigam and T Klindt, 'Industrie 4.0, das Internet der Dinge und das Recht' (2015) 16 *NJW* 1137 (hereafter Bräutigam and Klindt, 'Industrie 4.0'); T Kaufmann, *Geschäftsmodelle in Industrie 4.0 und dem Internet der Dinge* (2015); Schwab, *Die Vierte Industrielle Revolution* (2016); more reserved HJ Schlinkert, 'Industrie 4.0 – wie das Recht Schritt hält' (2017) 8 *ZRP* 222 *et seq*.

[2] Cf. A Börding and others, 'Neue Herausforderungen der Digitalisierung für das deutsche Zivilrecht' (2017) 2 *CR* 134; J Bormann, 'Die digitalisierte GmbH' (2017) 46 *ZGR* 621, 622; B Paal, 'Die digitalisierte GmbH' (2017) 46 *ZGR* 590, 592, 599 *et seq*.

[3] For digitalization of private law, see, e.g., K Langenbucher, 'Digitales Finanzwesen' (2018) 218 *AcP* 385 *et seq*.; G Teubner, 'Digitale Rechtssubjekte?' (2018) 218 *AcP* 155 *et seq*. (hereafter Teubner, 'Rechtssubjekte'); cf. further M Fries, 'PayPal Law und Legal Tech – Was macht die Digitalisierung mit dem Privatrecht?' (2016) 39 *NJW* 2860 *et seq* (hereafter Fries, 'Digitalisierung Privatrecht').

[4] For details, see, *e.g.*, H Eidenmüller, 'The Rise of Robots and the Law of Humans' (2017) 4 *ZEuP* 765 *et seq*.; G Spindler, 'Zukunft der Digitalisierung – Datenwirtschaft in der Unternehmenspraxis' (2018) 1–2 *DB* 41, 49 *et seq*. (hereafter Spindler, 'Zukunft').

[5] For details, see A Hildner and M Danzmann, 'Blockchain-Anwendungen für die Unternehmensfinanzierung' (2017) *CF* 385 *et seq*.; M Hüther and M Danzmann, 'Der Einfluss des Internet of Things und der Industrie 4.0 auf Kreditfinanzierungen' (2017) 15–16 *BB* 834 *et seq*.; R Nyffenegger and F Schär, 'Token Sales: Eine Analyse Des Blockchain-Basierten Unternehmensfinanzierungsinstruments' (2018) *CF* 121 *et seq*.; B Westermann, 'Daten als Kreditsicherheiten – eine Analyse des Datenwirtschaftsrechts de lege lata und de lege ferenda aus Sicht des Kreditsicherungsrechts' (2018) 26 *WM* 1205 *et seq*.

digital estate,[6] or online dispute resolution.[7] In fact, digital technologies challenge the entire national legal system including public and criminal law as well as EU and international law. Some even say we may face 'the beginning of the end for the law'.[8] In fact, this is not the end, but rather the time for a digital initiative. This chapter focuses on the changes that AI brings about in corporate law and corporate governance, especially in terms of the challenges for corporations and their executives.

From a conceptual perspective, AI applications will have a major impact on corporate law in general and corporate governance in particular. In practice, AI poses a tremendous challenge for corporations and their executives. As algorithms have already entered the boardroom, lawmakers must consider legally recognizing e-persons as directors and managers. The applicable law must deal with effects of AI on corporate duties of boards and their liabilities. The interdependencies of AI, delegation of leadership tasks, and the business judgement rule as a safe harbor for executives are of particular importance. A further issue to be addressed is how AI will change the decision-making process in corporations as a whole. This topic is closely connected with the board's duties in Big Data and Data Governance as well as the qualifications and responsibilities of directors and managers.

By referring to AI, I mean information technology systems that reproduce or approximate various cognitive abilities of humans.[9] In the same breath, we need to distinguish between strong AI and weak AI. Currently, strong AI does not exist.[10] There is no system really imitating a human being, such as a so-called superintelligence. Only weak AI is applied today. These are single technologies for smart human–machine interactions, such as machine learning or deep learning. Weak AI focuses on the solution of specific application problems based on the methods from math and computer science, whereby the systems are capable of self-optimization.[11]

[6] Cf. BGHZ 219, 243 (Bundesgerichtshof III ZR 183/17); A Kutscher, *Der digitale Nachlass* (2015); J Lieder and D Berneith, 'Digitaler Nachlass: Das Facebook-Urteil des BGH' (2018) 10 *FamRZ* 1486; C Budzikiewicz 'Digitaler Nachlass' (2018) 218 *AcP* 558 *et seq*.; H Ludgya, 'Digitales Update für das Erbrecht im BGB?' (2018) 1 *ZEV* 1 *et seq*.; C Sorge, 'Digitaler Nachlass als Knäuel von Rechtsverhältnissen' (2018) 6 *MMR* 372 *et seq*.; see also Deutscher Bundestag, 'Kleine Anfrage der Abgeordneten Roman Müller-Böhm et al. BT-Drucks. 19/3954' (2018); as to this J Lieder and D Berneith, 'Digitaler Nachlass – Sollte der Gesetzgeber tätig warden?' (2020) 3 *ZRP* 87 *et seq*.

[7] For details, see Fries, 'Digitalisierung Privatrecht' (n 3) 2681 *et seq*.; M Grupp 'Legal Tech – Impulse für Streitbeilegung und Rechtsdienstleistung' (2014) 8+9 *AnwBl*. 660 *et seq*.; J Wagner, 'Legal Tech und Legal Robots in Unternehmen und den sie beratenden Kanzleien' (2017) 17 *BB* 898, 900 (hereafter Wagner, 'Legal Tech').

[8] V Boehme-Neßler, 'Die Macht der Algorithmen und die Ohnmacht des Rechts' (2017) 42 *NJW* 3031.

[9] For definitions, see, *e.g.*, M Herberger, '"Künstliche Intelligenz" und Recht – Ein Orientierungsversuch' (2018) 39 *NJW* 2825 *et seq*.; C Schael, 'Künstliche Intelligenz in der modernen Gesellschaft: Bedeutung der "Künstlichen Intelligenz" für die Gesellschaft' (2018) 42 *DuD* 547 *et seq*.; J Armour and H Eidenmüller, 'Selbstfahrende Kapitalgesellschaften?' (2019) 2–3 *ZHR* 169, 172 *et seq*. (hereafter Armour and Eidenmüller, 'Kapitalgesellschaften'); F Graf von Westphalen, 'Definition der Künstlichen Intelligenz in der Kommissionsmitteilung COM (2020) 64 final – Auswirkungen auf das Vertragsrecht' (2020) 35 *BB* 1859 *et seq*. (hereafter Graf von Westphalen, 'Definition'); P Hacker, 'Europäische und nationale Regulierung von Künstlicher Intelligenz' (2020) 30 *NJW* 2142 *et seq*. (hereafter Hacker, 'Regulierung').

[10] Cf. U Noack, 'Organisationspflichten und -strukturen kraft Digitalisierung' (2019) 183 *ZHR* 105, 107 (hereafter Noack, 'Organisationspflichten'); U Noack, 'Der digitale Aufsichtsrat' in B Grunewald, J Koch, and J Tielmann (eds), *Festschrift für Eberhard Vetter* (2019) 497, 500 (hereafter Noack, 'Aufsichtsrat'); for a different use of this wording, see L Strohn, 'Die Rolle des Aufsichtsrats beim Einsatz von Künstlicher Intelligenz' (2018) 182 *ZHR* 371 *et seq*. (hereafter Strohn, 'Rolle').

[11] See Deutscher Bundestag, 'Antwort der Bundesregierung, Erarbeitung einer KI-Strategie der Bundesregierung, BT-Drucks. 19/5678' (2018) 2.

By referring to corporate governance, I mean a system by which companies are directed and controlled.[12] In continental European jurisdictions, such as Germany, a dual board structure is the prevailing system with a management board running the day-to-day business of the firm and a supervisory board monitoring the business decisions of the management board. In Anglo-American jurisdictions, such as the United States (US) and the United Kingdom (UK), the two functions of management and supervision are combined within one unitary board – the board of directors.[13]

II. ALGORITHMS AS DIRECTORS

The first question is, "Could and should algorithms act as directors?" In 2014, newspapers reported that a venture capital firm had just appointed an algorithm to its board of directors. The Hong Kong based VC firm Deep Knowledge Ventures was supposed to have appointed an algorithm called Vital (an abbreviation for Validating Investment Tool for Advancing Life Sciences) to serve as a director with full voting rights and full decision-making power over corporate measures.[14] In fact, Vital only had an observer and adviser status with regard to the board members, which are all natural persons.[15]

Under German law according to sections 76(3) and 100(1)(1) AktG,[16] the members of the management board and the supervisory board must be natural persons with full legal capacity. Not even corporations are allowed to serve as board members. That means, in order to appoint algorithms as directors, the law must be changed.[17] Actually, the lawmaker could legally recognize e-persons as directors. However, the lawmaker should not do so, because there is a reason for the exclusion of legal persons and algorithms under German law. Both lack personal liability and personal accountability for the management and the supervision of the company.[18]

[12] Cf. UH Schneider and C Strenger, *Die "Corporate Governance-Grundsätze" der Grundsatzkommission Corporate Governance (German Panel on Corporate Governance)* (2000) 106, 107; R Marsch-Barner, '§ 2 Corporate Governance marginal number 2.1' in R Marsch-Barner and F Schäfer (eds), *Handbuch börsennotierte AG* (4th ed. 2018); J Koch, '§ 76 margin number 37' in U Hüffer and J Koch (eds) *Aktiengesetz* (14th ed. 2020); HJ Böcking and L Bundle, '§ 2 marginal number 6' in KJ Hopt, JH Binder, and HJ Böcking (eds), *Handbuch Corporate Governance von Banken und Versicherungen* (2nd ed. 2020); A v Werder, 'DCGK Präambel marginal number 10' in T Kremer and others (eds), *Deutscher Corporate Governance Kodex* (8th ed. 2021).

[13] For a comparative overview, see J Lieder, *'Der Aufsichtsrat im Wandel der Zeit'* (2006) 636 *et seq*. (hereafter Lieder 'Aufsichtsrat').

[14] R Wile, 'A Venture Capital Firm Just Named an Algorithm to Its Board of Directors' (*Business Insider*, 13 May 2014) www.businessinsider.com/vital-named-to-board-2014-5?r=US&IR=T.

[15] See N Burridge, 'Artificial Intelligence gets a seat in the boardroom' (*Nikkei Asia*, 10 May 2017) https://asia.nikkei.com/Business/Artificial-intelligence-gets-a-seat-in-the-boardroom.

[16] *Aktiengesetz (AktG)* = Stock Corporation Act of 6 September 1965, Federal Law Gazette I, 1089. For the English version that has been used in this paper, see Rittler, *German Corporate Law* (2016) as well as Norton Rose Fullbright, 'German Stock Coroporation Act (*Aktiengesetz*)' (*Norton Rose Fullbright*, 10 May 2016) www.nortonrosefulbright.com/-/media/files/nrf/nrfweb/imported/german-stock-corporation-act.pdf.

[17] Cf. H Fleischer, '§ 93 marginal number 129' in M Henssler (ed), *BeckOGK Aktiengesetz* (15 January 2020); F Möslein, 'Digitalisierung im Gesellschaftsrecht: Unternehmensleitung durch Algorithmen und künstliche Intelligenz?' (2018) 5 *ZIP* 204, 207 *et seq*. (hereafter Möslein, 'Digitalisierung'); Strohn, 'Rolle'(n 10) 371; R Weber, A Kiefner, and S Jobst, 'Künstliche Intelligenz und Unternehmensführung' (2018) 29 NZG 1131 (1136) (hereafter 'Weber, Kiefner, and Jobst 'Unternehmensführung'); see further H Fleischer, 'Algorithmen im Aufsichtsrat' (2018) 9 *Der Aufsichtsrat* 121 (hereafter Fleischer, 'Algorithmen'); A Sattler, 'Der Einfluss der Digitalisierung auf das Gesellschaftsrecht' (2018) 39 *BB* 2243, 2248 (hereafter Sattler, 'Einfluss'); Wagner, 'Legal Tech' (n 7) 1098.

[18] See B Kropff, *Begründung zum Regierungsentwurf zum Aktiengesetz 1965* (1965) 135: 'Der Entwurf gestattet es nicht, juristische Personen zu wählen, weil die Überwachungspflicht die persönliche Tätigkeit einer verantwortlichen Person voraussetzt.' Cf. further Lieder, 'Aufsichtsrat' (n 13) 367 *et seq*.

Nevertheless, the European Parliament enacted a resolution with recommendations to the Commission on Civil Law Rules on Robotics, and suggested therein

> creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently.[19]

The most fundamental requirement for legally recognizing an e-person would be its own liability – either based on an ownership fund or based on a mandatory liability insurance. In case corporations are appointing AI entities as directors (or apply it otherwise), they should be strictly liable for damages caused by AI applications in order to mitigate the particular challenges and potential risks of AI.[20] This is because strict liability would not only delegate the risk assessment and thus control the level of care and activity, but would also create an incentive for further developing this technology.[21] At the same time, creditors of the company should be protected by a compulsory liability insurance, whereas piercing the corporate veil, that is, a personal liability of the shareholders, must remain a rare exception.[22] However, at an international level, regulatory competition makes it difficult to guarantee comparable standards. Harmonization can only be expected (if ever) in supranational legal systems, such as the European Union.[23] In this context, it is noteworthy that the EU Commission's White Paper on AI presented in 2020 does not address the questions of the legal status of algorithms at all.[24]

However, even if we were to establish such a liability safeguard, there is no self-interested action of an algorithm as long as there is no strong AI. True, circumstances may change in the future due to technological progress. However, there is a long and winding road to the notorious superintelligence.[25] Conversely, weak AI only carries out actions in the third-party interest of people or organizations, and is currently not in a position to make its own value decisions and judgemental considerations.[26] In the end, current algorithms are nothing more than digital slaves, albeit slaves with superhuman abilities. In addition, the currently applicable incentive

[19] M Delvaux, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL))' (*European Parliament*, 27 January 2015) www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html; see, *e.g.*, MF Lohmann, 'Ein europäisches Roboterrecht – überfällig oder überflüssig?' (2017) 6 *ZRP* 168; R Schaub, 'Interaktion von Mensch und Maschine' (2017) 7 *JZ* 342, 345 *et seq*.; J-E Schirmer, 'Rechtsfähige Roboter?' (2016) 13 *JZ* 660 *et seq*.; J Taeger, 'Die Entwicklung des IT-Rechts im Jahr 2016' (2016) 52 *NJW* 3764.

[20] Cf. Bräutigam and Klindt, 'Industrie 4.0' (n 1) 1138; Teubner, 'Rechtssubjekte' (n 3) 184; DA Zetzsche, 'Corporate Technologies – Zur Digitalisierung im Aktienrecht' (2019) 1-02 AG 1 (10) (hereafter Zetzsche, 'Technologies').

[21] Cf. G Borges, 'Haftung für selbstfahrende Autos' (2016) 4 CR 272, 278; H Kötz and G Wagner, *Deliktsrecht* (13th ed. 2016) marginal number 72 *et seq*.; H Zech, 'Künstliche Intelligenz und Haftungsfragen' (2019) 2 *ZfPW* 198, 214.

[22] Cf. Armour and Eidenmüller, 'Kapitalgesellschaften' (n 9) 185 *et seq*.

[23] Ibid, 186 *et seq*.

[24] European Commission, 'White Paper On Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final' (*EUR-Lex*, 19 February 2020) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2020:65:FIN; see Graf von Westphalen, 'Definition' (n 9) 1859 *et seq*.; Hacker, 'Regulierung' (n 9), 2142 *et seq*.

[25] Cf. further B Schölkopf, 'Der Mann, der den Computern das Lernen beibringt' *Frankfurter Allgemeine Zeitung* (26 February 2020): "Wir sind extrem weit davon entfernt, dass seine Maschine intelligenter ist als ein Mensch."; L Enriques and DA Zetzsche, 'Corporate Technologies and the Tech Nirvana Fallacy' (2019) ECGI Law Working Paper N° 457/2019, 58 https://ecgi.global/sites/default/files/working_papers/documents/finalenriqueszetzsche.pdf (hereafter Enriques and Zetzsche, 'Corporate Technologies'): "Only if and when humans relinquish corporate control to machines, may the problem at the core of corporate governance be solved; but by then humans will have more pressing issues to worry about than corporate governance."

[26] Cf. further P Krug, *Haftung im Rahmen der Anwendung von künstlicher Intelligenz: Betrachtung unter Berücksichtigung der Besonderheiten des steuerberatenden Berufsstandes* (2020) 74, 76; Möslein, 'Digitalisierung' (n 17) 207; Noack, 'Aufsichtsrat' (n 10) 506.

system of corporate law and governance would have to be adapted to AI directors, because – unlike human directors – duties of loyalty can hardly be applied to them, but rather they decide according to algorithmic models.[27] At present, only humans have original creative power, only they are capable of making decisions and acting in the true sense of the word.[28]

III. MANAGEMENT BOARD

Given the current limitations of AI, we will continue to have to get by with human directors for the next few decades. Although algorithms do not currently appear suitable for making independent corporate decisions, AI can nonetheless support human directors in their management and monitoring tasks. AI is already used in practice to analyze and forecast the financial development of a company, but also to identify the need for optimization in an entrepreneurial value chain.[29] In addition, AI applications are used in the run-up to mergers and acquisitions (M&A) transactions,[30] namely as part of due diligence, in order to simplify particularly labor-intensive processes when checking documents. Algorithms are also able to recognize unusual contract clauses and to summarize essential parameters of contracts, even to create contract templates themselves.[31] Further examples for the use of AI applications are cybersecurity[32] and compliance management systems.[33]

1. *Legal Framework*

With regard to the German corporate governance system, the management board is responsible for running the company.[34] Consequently, the management board also decides on the overall corporate strategy, the degree of digitalization and the use of AI applications.[35] The supervisory board monitors the business decisions of the management board, decides on the approval of particularly important measures,[36] as well as on the appointment and removal of the management board members;[37] whereas the shareholders meeting does not determine a company's digitalization structures.[38]

2. *AI Related Duties*

In principle, the use of AI neither constitutes a violation of corporate law or the articles of association,[39] nor is it an expression of bad corporate governance. Even if the use of AI is associated

[27] Möslein, 'Digitalisierung' (n 17) 206.
[28] Cf. M Auer, 'Der Algorithmus kennt keine Moral' *Frankfurter Allgemeine Zeitung* (29 April 2020).
[29] On this and the following, see Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1131.
[30] Cf. Noack, 'Organisationspflichten' (n 10) 119.
[31] Cf. M Grub and S Krispenz, 'Auswirkungen der Digitalisierung auf M&A Transaktionen' (2018) 5 *BB* 235, 238; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1131.
[32] For details, see Noack, 'Organisationspflichten' (n 10) 124 *et seq*.
[33] Cf. Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1131; Noack, 'Organisationspflichten' (n 10) 132 *et seq*.; Zetzsche, 'Technologies' (n 20) 5.
[34] Section 76(1) AktG.
[35] Cf. Noack, 'Organisationspflichten' (n 10) 115 *et seq*.
[36] Section 111(4) AktG.
[37] Section 84 AktG.
[38] Cf. Section 119 AktG.
[39] Cf. Strohn, 'Rolle' (n 10) 371, 376.

with risks, it is difficult to advise companies – as the safest option – to forego it completely.[40] Instead, the use of AI places special demands on the management board members.

a. General Responsibilities

Managers must have a fundamental understanding of the relevant AI applications, of their potentials, suitability, and risks. However, the board members do not need to have in-depth knowledge about the detailed functioning of a certain AI application. In particular, the knowledge of an IT expert cannot be demanded, nor a detailed examination of the material correctness of the decision.[41] Rather, they need to have an understanding of the scope and limits of an application and possible results and outcomes of the application in order to perform plausibility checks to prevent incorrect decisions quickly and effectively.[42] The management board has to ensure, through test runs, the functionality of the application with regard to the concrete fulfilment of tasks in the specific company environment.[43] If, according to the specific nature of the AI application, there is the possibility of an adjustment to the concrete circumstances of the company, for example, with regard to the firm's risk profile or statutory provisions, then the management board is obliged to carry out such an adjustment.[44] During the use of the AI, the board of directors must continuously evaluate and monitor the working methods, information procurement, and information evaluation as well as the results achieved.

The management board must implement a system that eliminates, as far as possible, the risks and false results that arise from the use of AI. This system must assure that anyone who uses AI knows the respective scope of possible results of an application so that it can be determined whether a concrete result is still within the possible range of results. However, that can hardly be determined abstractly, but requires a close look at the concrete AI application. Furthermore, the market standard is to be included in the analysis. If all companies in a certain industry use certain AI applications that are considered safe and effective, then an application by other companies will rarely prove to breach a management board's duty of care.

Under these conditions, the management board is allowed to delegate decisions and tasks to an AI application.[45] This is not contradicted by the fact that algorithms lack legal capacity, because in this context the board's own duties are decisive.[46] In any event, a blanket self-commitment to the results of an AI application is incompatible with the management responsibility and personal accountability of the board members.[47] At all times, the applied AI must be manageable and controllable in order to ensure that no human loss of control occurs and the

[40] But see Strohn, 'Rolle' (n 10) 376; rightly contested by Noack, 'Aufsichtsrat' (n 10) 502.

[41] Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1133; cf. further J Wagner, 'Legal Tech und Legal Robots in Unternehmen und den sie beratenden Kanzleien – Teil 2: Folgen für die Pflichten von Vorstandsmitgliedern bzw. Geschäftsführern und Aufsichtsräten' (2018) 20 *BB* 1097, 1099 (hereafter Wagner, 'Legal Tech 2'); Möslein, 'Digitalisierung' (n 17) 208 *et seq*.

[42] Cf. further Sattler, 'Einfluss' (n 17) 2248.

[43] M Becker and P Pordzik, 'Digitale Unternehmensführung' (2020) 3 *ZfPW* 334, 349 (hereafter Becker and Pordzik, 'Unternehmensführung').

[44] On this and the following, see Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

[45] Becker and Pordzik, 'Unternehmensführung' (n 43) 344; Noack, 'Organisationspflichten' (n 10) 117; O Lücke, 'Der Einsatz von KI in der und durch die Unternehmensführung' (2019) 35 *BB* 1986, 1989, and 1992 (hereafter Lücke, 'KI'); for a different view, see V Hoch, 'Anwendung Künstlicher Intelligenz zur Beurteilung von Rechtsfragen im unternehmerischen Bereich' (2019) 219 *AcP* 648, 672.

[46] Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

[47] Möslein, 'Digitalisierung' (n 17) 208 *et seq*.; Wagner, 'Legal Tech 2' (n 41) 1098 *et seq*., 1101; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

decision-making process is comprehensible. The person responsible for applying AI in a certain corporate setting must always be able to operate the off-switch. In normative terms, this requirement is derived from section 91(2) AktG, which obliges the management board to take suitable measures to identify, at an early stage, developments that could jeopardize the continued existence of the company.[48] In addition, the application must be protected against external attacks, and emergency precautions must be implemented in the event of a technical malfunction.[49]

b. Delegation of Responsibility

The board may delegate the responsibility for applying AI to subordinate employees, but it is required to carefully select, instruct, and supervise the delegate.[50] Under the prevailing view, core tasks, however, cannot be delegated, as board members are not allowed to evade their leadership responsibility.[51] Such non-delegable management tasks of the management board include basic measures with regard to the strategic direction, business policy and the organization of the company.[52] The decision as to whether and to what extent AI should be used in the company is also a management measure that cannot be delegated under the prevailing view.[53] Only the preparation of decisions by auxiliary persons is permissible, as long as the board of directors makes the decision personally and of its own responsibility. In this respect, the board is responsible for the selection of AI use and the application of AI in general. The board has to provide the necessary information, must exclude conflicts of interest and has to perform plausibility checks of the results obtained. Furthermore, the managers must conduct an ongoing monitoring and ensure that the assigned tasks are properly performed.

c. Data Governance

AI relies on extensive data sets (Big Data). In this respect, the management board is responsible for a wide scope and high quality of the available data, for the suitability and training of AI applications, and for the coordination of the model predictions with the objectives of the respective company.[54] In addition, the board of directors must observe data protection

[48] Cf. Zetzsche, 'Technologies' (n 20) 7.

[49] Becker and Pordzik, 'Digitale Unternehmensführung' (n 43) 352; D Linardatos, 'Künstliche Intelligenz und Verantwortung' (2019) 11 *ZIP* 504, 508; Lücke, 'KI' (n 45) 1993.

[50] M Dreher, 'Nicht delegierbare Geschäftsleiterpflichten' in S Grundmann and others (eds), *Festschrift für Klaus J. Hopt zum 70. Geburtstag* (2010) 517, 536; H Fleischer, '§ 93 marginal number 98 *et seq.*' in M Henssler (ed), *BeckOGK Aktiengesetz* (15 January 2020); HC Grigoleit and L Tomasic, '§ 93 marginal number 38' in HC Grigoleit, *Aktiengesetz* (2020); Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

[51] Cf. M Dreher, 'Nicht delegierbare Geschäftsleiterpflichten' in S Grundmann and others (eds) *Festschrift für Klaus J Hopt zum 70. Geburtstag* (2010) 517, 527; with a specific focus on AI use, see Möslein, 'Digitalisierung' (n 17) 208 *et seq.*; Wagner, 'Legal Tech 2' (n 41) 1098; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

[52] HJ Mertens and A Cahn, '§ 76 marginal number 4' in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010); M Weber, '§ 76 marginal number 8' in W Hölters (ed), *AktG* (3rd ed. 2017); Weber, Kiefner and Jobst, 'Unternehmensführung' (n 17) 1132.

[53] M Kort, '§ 76 marginal number 37' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); G Spindler, 'Haftung der Geschäftsführung für IT-Sachverhalte' (2017) 11 *CR* 715, 722; G Spindler, 'Gesellschaftsrecht und Digitalisierung' (2018) 47 *ZGR* 17, 40 *et seq.* (hereafter Spindler, 'Gesellschaftsrecht'); Spindler, 'Zukunft' (n 4) 44; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

[54] For details, see Armour and Eidenmüller, 'Kapitalgesellschaften' (n 9) 176 *et seq.*; Krug, 'Haftung' (n 26) 78 *et seq.*; cf. further Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132 *et seq.*

law limits[55] and must pursue a non-discriminatory procedure.[56] If AI use is not in line with these regulations or other mandatory provisions, the management board violates the duty of legality.[57] In this case, the management board does not benefit from the liability privilege of the business judgement rule.[58]

Apart from that, the management board has an entrepreneurial discretion with regard to the proper organization of the company's internal knowledge organization.[59] The starting point is the management board's duty to ensure a legal, statutory, and appropriate organizational structure.[60] The specific scope and content of the obligation to organize knowledge depends largely on the type, size, and industry of the company and its resources.[61] However, if, according to these principles, there is a breach of the obligation to store, forward, and actually query information, then the company will be considered to have acted with knowledge or negligent ignorance under German law.[62]

d. Management Liability

If managers violate these obligations (and do not benefit from the liability privilege of the business judgement rule)[63], they can be held liable for damages to the company.[64] This applies in particular in the event of an inadmissible or inadequate delegation.[65] In order to mitigate the liability risk for management board members, they have to ensure that the whole framework of AI usage in terms of specific applications, competences, and responsibilities as well as the AI-related flow of information within the company is well designed and documented in detail. Conversely, board members are not liable for individual algorithmic errors as long as (1) the algorithm works reliably, (2) the algorithm does not make unlawful decisions, (3) there are no conflicts of interest, and (4) the AI's functioning is fundamentally overseen and properly documented.[66]

[55] For details, see T Hoeren and M Niehoff, 'KI und Datenschutz – Begründungserfordernisse automatisierter Entscheidungen' (2018) 1 *RW* 47 *et seq*.; cf. further CS Conrad, 'Kann die Künstliche Intelligenz den Menschen entschlüsseln? – Neue Forderungen zum Datenschutz: Eine datenschutzrechtliche Betrachtung der "Künstlichen Intelligenz"' (2018) 42 *DuD* 541 *et seq*.; M Rost, 'Künstliche Intelligenz: Normative und operative Anforderungen des Datenschutzes' (2018) 42 *DuD* 558.

[56] As to new types of discrimination risks, see JA Kroll and others, 'Accountable Algorithms' (2017) 165 *U Pa L Rev* 633, 679 *et seq*.; B Paal, 'Vielfaltsicherung im Suchmaschinensektor' (2015) 2 *ZRP* 34, 35; H Steege, 'Algorithmenbasierte Diskriminierung durch Einsatz von Künstlicher Intelligenz: Rechtsvergleichende Überlegungen und relevante Einsatzgebiete' (2019) 11 *MMR* 715 *et seq*.

[57] Cf. F König, 'Haftung für Cyberschäden: Auswirkungen des neuen Europäischen Datenschutzrechts auf die Haftung von Aktiengesellschaften und ihrer Vorstände' (2017) 8 *AG* 262, 268 *et seq*.; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1135.

[58] See infra Section III 3.

[59] G Spindler and A Seidel, 'Die zivilrechtlichen Konsequenzen von Big Data für Wissenszurechnung und Aufklärungspflichten' (2018) 30 *NJW* 2153, 2154 (hereafter Spindler and Seidel, 'Big Data'); G Spindler and A Seidel, 'Wissenszurechnung und Digitalisierung' in G Spindler and others (eds), *Unternehmen, Kapitalmarkt, Finanzierung. Festschrift für Reinhard Marsch-Barner* (2018) 549, 552 *et seq*.

[60] Cf. BGHZ 132, 30 (37) (Bundesgerichtshof V ZR 239/94); P Hemeling, 'Organisationspflichten des Vorstands zwischen Rechtspflicht und Opportunität' (2011) 175 *ZHR* 368, 380.

[61] Spindler and Seidel, 'Big Data' (n 59) 2154.

[62] Cf. HC Grigoleit, 'Zivilrechtliche Grundlagen der Wissenszurechnung' (2017) 181 *ZHR* 160 *et seq*.; M Habersack and M Foerster, '§ 78 marginal number 39' in H Hirte, P O Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); Sattler, 'Einfluss' (n 17) 2248; Spindler, 'Wissenszurechnung in der GmbH, der AG und im Konzern'(2017) 181 ZHR 311 *et seq*.

[63] See infra Section III 3.

[64] AktG, section 93(2).

[65] Cf. further Möslein, 'Digitalisierung' (n 17) 210 *et seq*.

[66] Cf. Möslein, 'Digitalisierung' (n 17) 211.

Comprehensive documentation of the circumstances that prompted the management board to use a certain AI and the specific circumstances of its application reduces the risk of being sued for damages by the company. This ensures, in particular, that the members of the management board can handle the burden of proof incumbent on them according to section 93 (2)(2) AktG. They will achieve this better, the more detailed the decision-making process regarding the use of AI can be understood from the written documents.[67] This kind of documentation by the management board is to be distinguished from general documentation requirements discussed at the European and national level for the development of AI models and for access authorization to this documentation, the details of which are beyond the scope of this chapter.[68]

e. Composition of the Management Board

In order to cope with the challenges that the use of AI applications causes, the structure and composition of the management and the board has already changed significantly. That manifests itself in the establishment of new management positions, such as a Chief Information Officer (CIO)[69] or a Chief Digital Officer (CDO).[70] Almost half of the 40 largest German companies have such a position at board level.[71]

In addition, soft factors are becoming increasingly important in corporate management. Just think of the damage to the company's reputation, which is one of the tangible economic factors of a company today.[72] Under the term Corporate Digital Responsibility (CDR), specific responsibilities are developing for the use of AI and other digital innovations.[73] For example, Deutsche Telekom AG has enacted nine guidelines for responsible AI in a corporate setting. SAP SE established an advisory board for responsible AI consisting of experts from academia, politics, and the industry. These developments, of course, have an important influence on the overall knowledge attribution within the company and a corporate group. AI and Big Data make information available faster and facilitate the decision-making process at board level. Therefore, the management board must examine whether the absence of any AI application in the information gathering and decision-making process is in the best interest of a company. However, a duty to use AI applications only exists in exceptional cases and depends on the market standard in the respective industry. The greater the amount of data to be managed and the more complex and calculation-extensive the decisions in question, the more likely it is that the management board will be obliged to use AI.[74]

[67] Cf. Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1136; in general, see W Hölters, '§ 93 marginal number 36' in W Hölters (ed), *AktG* (3rd ed. 2017); HJ Mertens and A Cahn, '§ 93 marginal number 36' in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010); G Spindler, '§ 93 marginal number 58' in W Goette and M Habersack (eds) '*Münchener Kommentar zum AktG*' (5th ed. 2019).

[68] Hacker, 'Regulierung' (n 9) 2143 *et seq*.

[69] Cf. Sattler, 'Einfluss' (n 17) 2248.

[70] Cf. M Kaspar, 'Aufsichtsrat und Digitalisierung' (2018) *BOARD* 202, 203.

[71] Cf. Noack, 'Aufsichtsrat' (n 10) 502 *et seq*.

[72] For details, see U Schmolke and L Klöhn, 'Unternehmensreputation (Corporate Reputation)' (2015) 18 *NZG* 689 *et seq*.

[73] On this and the following, see Noack, 'Organisationspflichten' (n 10) 112 *et seq*.; Noack, 'Aufsichtsrat' (n 10) 503 *et seq*.; cf. further F Möslein, 'Corporate Digital Responsibility' in S Grundmann and others (eds), *Festschrift für Klaus J Hopt zum 80. Geburtstag* (2020) 805 *et seq*.

[74] Möslein, 'Digitalisierung' (n 17) 209.

3. *Business Judgement Rule*

This point is closely connected with the application of the business judgement rule as a safe harbour for AI use. Under the general concept of the business judgement rule that is well-known in many jurisdictions[75], as it is in Germany according to section 93(1)(2) AktG, a director cannot be held liable for an entrepreneurial decision if there is no conflict of interest and she had good reason to assume she was acting based on adequate information and for the benefit of the company.

a. Adequate Information

The requirement of adequate information depends significantly on the ability to gather and analyse information. Taking into account all the circumstances of the specific individual case, the board of directors has a considerable amount of leeway to judge which information is to be obtained from an economic point of view in the time available and to be included in the decision-making process. Neither a comprehensive nor the best possible, but only an appropriate information basis is necessary.[76] In addition, the appropriateness is to be assessed from the subjective perspective of the board members ('could reasonably assume'), so that a court is effectively prevented during the subsequent review from substituting its own understanding of appropriateness for the subjective assessment of the decision-maker.[77] In the context of litigation, a plausibility check based on justifiability is decisive.[78]

In general, the type, size, purpose, and organization of the company as well as the availability of a functional AI and the data required for operation are relevant for answering the question of the extent to which AI must be used in the context of the decision-making preparation based on information. The cost of the AI system and the proportionality of the information procurement must also be taken into account.[79] If there is a great amount of data to be managed and a complex and calculation-intensive decision to be made, AI and Big Data applications are of major importance and the members of the management board will hardly be able to justify not using AI.[80] Conversely, the use of AI to obtain information is definitely not objectionable.[81]

[75] For a comparative view, see H Merkt, 'Rechtliche Grundlagen der Business Judgment Rule im internationalen Vergleich zwischen Divergenz und Konvergenz' (2017) 46 *ZGR* 129 *et seq*.

[76] For details, see J Lieder, 'Unternehmerische Entscheidungen des Aufsichtsrats' (2018) 47 *ZGR* 523, 555 (hereafter Lieder, 'Entscheidungen').

[77] Cf. further KJ Hopt and M Roth, '§ 93 marginal number 102' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); H Fleischer, '§ 93 marginal number 90' in M Henssler (ed), *BeckOGK, Aktiengesetz* (15 January 2020); M Kock and R Dinkel, 'Die zivilrechtliche Haftung von Vorständen für unternehmerische Entscheidungen – Die geplante Kodifizierung der Business Judgment Rule im Gesetz zur Unternehmensintegrität und Modernisierung des Anfechtungsrechts' (2004) 10 *NZG* 441, 444; Lieder, 'Entscheidungen' (n 76) 557; for a different view, see W Goette, 'Gesellschaftsrechtliche Grundfragen im Spiegel der Rechtsprechung' (2008) 37 *ZGR* 436, 447 *et seq*.: purely objective approach.

[78] Cf. H Fleischer in M Henssler (ed), *BeckOGK, Aktiengesetz* (15 January 2020) § 93 marginal number 91; J Koch in U Hüffer and J Koch (eds) *Aktiengesetz* (14th ed. 2020) § 93 marginal number 21; HJ Mertens and A Cahn in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010) § 93 marginal number 34; Lieder, 'Entscheidungen' (n 76) 557; J Redeke, 'Zur gerichtlichen Kontrolle der Angemessenheit der Informationsgrundlage im Rahmen der Business Judgement Rule nach § 93 Abs. 1 S. 2 AktG' (2011) 2 *ZIP* 59, 60 *et seq*.

[79] Cf. Noack, 'Organisationspflichten' (n 10) 122.

[80] Cf. Becker and Pordzik, 'Unternehmensführung' (n 43) 347; Möslein, 'Digitalisierung' (n 17) 209 *et seq*., 212; Sattler, 'Einfluss' (n 17) 2248; Spindler, 'Gesellschaftsrecht' (n 53) 43; Spindler, 'Zukunft' (n 4) 45; Weber, Kiefner, and Jobst (n 17) 1134.

[81] Wagner, 'Legal Tech 2' (n 41) 1100; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1132.

b. Benefit of the Company

Furthermore, the board of directors must reasonably assume to act in the best interest of the company when using AI. This criterion is to be assessed from an *ex ante* perspective, not *ex post*.[82] According to the mixed-subjective standard, it depends largely on the concrete perception of the acting board members at the time of the entrepreneurial decision.[83] In principle, the board of directors is free to organize the operation of the company according to its own ideas, as long as it stays within the limits of the best interest of the corporation[84] that are informed solely by the existence and the long-term and sustainable profitability of the company.[85] Only when the board members act in a grossly negligent manner or take irresponsible risks do they act outside the company's best interest.[86] Taking all these aspects into account, the criterion of acceptability proves to be a suitable benchmark.[87]

In the specific decision-making process, all advantages and disadvantages of using or delegating the decision to use AI applications must be included and carefully weighed against one another for the benefit of the company. In this context, however, it cannot simply be seen as unacceptable and contrary to the welfare of the company that the decisions made by or with the support of AI can no longer be understood from a purely human perspective.[88] On the one hand, human decisions that require a certain originality and creativity cannot always be traced down to the last detail. On the other hand, one of the major potentials of AI is to harness particularly creative and original ideas in the area of corporate management. AI can, therefore, be used as long as its use is not associated with unacceptable risks. The business judgement rule allows the management board to consciously take at least justifiable risks in the best interest of the company.

[82] Cf. Deutscher Bundestag, 'Begründung zum Regierungsentwurf, BT-Drucks. 15/5092' (2005) 11; T Bürgers, '§ 93 marginal number 15' in T Bürgers and T Körber (eds), *AktG* (4th ed. 2017); B Dauner-Lieb, '§ 93 AktG marginal number 23' in M Henssler and L Strohn (eds), *Gesellschaftsrecht* (5th ed. 2021); KJ Hopt and M Roth, '§ 93 marginal number 101' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); W Hölters, '§ 93 marginal number 39' in W Hölters (ed), *AktG* (3rd ed. 2017); HJ Mertens and A Cahn, '§ 93 marginal number 23' in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010); Lieder, 'Entscheidungen' (n 76) 577.

[83] Similarly H Fleischer, '§ 93 marginal number 92' in M Henssler (ed), *BeckOGK, Aktiengesetz* (15 January 2020); KJ Hopt and M Roth, '§ 93 marginal number 101' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2016); J Koch, '§ 93 marginal number 21' in U Hüffer and J Koch (eds), *Aktiengesetz* (14th ed. 2020); Lieder, 'Entscheidungen' (n 76) 577; for a different opinion (objective standard): W Hölters, '§ 93 marginal number 39' in W Hölters, *AktG* (3rd ed. 2017); HJ Mertens and A Cahn in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010) § 93 marginal number 23.

[84] Cf. H Fleischer, '§ 76 marginal number 27' in M Henssler (ed), *BeckOGK, Aktiengesetz* (15 January 2020); M Kort, '§ 76 marginal number 60' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG*, (5th ed. 2015); G Spindler, '§ 76 marginal number 67 ff.' in W Goette and M Habersack (eds) '*Münchener Kommentar zum AktG*' (5th ed. 2019); P Ulmer, 'Aktienrecht im Wandel' (2002) 202 *AcP* 143, 158 *et seq*.

[85] Cf. OLG Hamm AG 1995, 512, 514; B Dauner-Lieb, '§ 93 AktG marginal number 23' in M Henssler and L Strohn (eds), *Gesellschaftsrecht*, (5th ed. 2021); J Koch, '§ 76 marginal number 34' in U Hüffer and J Koch (eds) *Aktiengesetz* (14th ed. 2020); M Kort, '§ 76 marginal number 52' in H Hirte, P O Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); HJ Mertens and A Cahn, '§ 76 marginal number 17' in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010) 22; Lieder, 'Entscheidungen' (n 76) 577–578.

[86] Vgl. BGHZ 135, 244 (253–254) (Bundesgerichtshof II ZR 175/95); H Fleischer, '§ 93 marginal number 99' in M Henssler (ed), *BeckOGK, Aktiengesetz* (15 January 2020); J Koch, '§ 93 marginal number 23' in U Hüffer and J Koch (eds), *Aktiengesetz* (14th ed. 2020).

[87] Cf. T Drygala, '§ 116 marginal number 15' in K Schmidt and M Lutter (eds), *AktG* (4th ed. 2020); J Koch, '§ 93 marginal number 23' in U Hüffer and J Koch (eds), *Aktiengesetz* (14th ed. 2020); Lieder, 'Entscheidungen' (n 76) 578 with examples.

[88] But see Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1135.

However, the management board may also conclude that applying AI is just too much of a risk for the existence or the profitability of the firm and therefore may refrain from it without taking a liability risk under section 93(1)(2) AktG.[89] The prerequisite for this is that the board performs a conscious act of decision-making.[90] Otherwise, acting in good faith for the benefit of the company is ruled out *a priori*. This decision can also consist of a conscious toleration or omission.[91] The same applies to intuitive action,[92] even if in this case the other requirements of section 93(1)(2) AktG must be subjected to a particularly thorough examination.[93] Furthermore, in addition to the action taken, there must have been another alternative,[94] even if only to omit the action taken. Even if the decision makers submit themselves to an actual or supposed necessity,[95] they could at least hypothetically have omitted the action. Apart from that, the decision does not need to manifest itself in a formal act of forming a will; in particular, a resolution by the collective body is not a prerequisite. Conversely, with a view to a later (judicial) dispute, it makes sense to sufficiently document the decision.[96]

c. Freedom from Conflicts of Interest

The executive board must make the decision for or against the use of AI free of extraneous influences and special interests.[97] The business judgement rule does not apply if the board members are not solely guided by the points mentioned above, but rather pursue other, namely self-interested goals. If the use of AI is not based on inappropriate interests and the board of directors has not influenced the parameters specified for the AI in a self-interested manner, the use

[89] Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1134.

[90] Deutscher Bundestag, 'Begründung zum Regierungsentwurf, BT-Drucks. 15/5092' (2005) 11; B Dauner-Lieb, '§ 93 AktG marginal number 20' in M Henssler and L Strohn (eds), *Gesellschaftsrecht* (5th ed. 2021); J Koch, '§ 93 marginal number 16' in U Hüffer and J Koch (eds), *Aktiengesetz* (14th ed. 2020); Lieder, 'Entscheidungen' (n 76) 532.

[91] Deutscher Bundestag, 'Begründung zum Regierungsentwurf, BT-Drucks. 15/5092' (2005) 11; T Bürgers '§ 93 marginal number 15' in T Bürgers and T Körber (eds), *AktG* (4th ed. 2017); H Fleischer, '§ 93 marginal number 97' in M Henssler (ed), *BeckOGK, Aktiengesetz* (15 January 2020); HC Ihrig, 'Reformbedarf beim Haftungstatbestand des § 93 AktG' (2004) 43 *WM* 2098, 2105.

[92] KJ Hopt and M Roth, '§ 93 marginal number 80' in H Hirte, P O Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); for a different view, see G Spindler, '§ 93 marginal number 51' in W Goette and M Habersack (eds), *'Münchener Kommentar zum AktG'* (5th ed. 2019).

[93] Lieder, 'Entscheidungen' (n 76) 532; too indiscriminately negative, however, G Spindler, '§ 93 marginal number 51' in W Goette and M Habersack (eds) *'Münchener Kommentar zum AktG'* (5th ed. 2019); H Hamann, 'Reflektierte Optimierung oder bloße Intuition?' (2012) *ZGR* 817, 825 *et seq*.

[94] T Bürgers, '§ 93 marginal number 11' in T Bürgers and T Körber, *AktG* (4th ed. 2017); B Dauner-Lieb, '§ 93 AktG marginal number 20' in M Henssler and L Strohn (eds), *Gesellschaftsrecht* (5th ed. 2021); M Graumann, 'Der Entscheidungsbegriff in § 93 Abs 1 Satz 2 AktG' (2011) *ZGR* 293, 296; for a different view, see KJ Hopt and M Roth, '§ 93 marginal number 80' in H Hirte, P O Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015).

[95] Cf. KJ Hopt and M Roth, '§ 93 marginal number 80' in H Hirte, P O Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015).

[96] J Koch, '§ 93 marginal numbers 16, 22' in U Hüffer and J Koch (eds), *Aktiengesetz* (14th ed. 2020); G Krieger and V Sailer-Coceani, '§ 93 marginal number 41' in K Schmidt and M Lutter (eds), *AktG* (4th ed. 2020); M Lutter, 'Die Business Judgment Rule und ihre praktische Anwendung' (2007) 18 *ZIP* 841, 847; Lieder, 'Entscheidungen' (n 76) 533.

[97] Deutscher Bundestag, 'Begründung zum Regierungsentwurf, BT-Drucks. 15/5092' (2005) 11; *BGHZ* 135, 244 (253) (Bundesgerichtshof II ZR 175/95); B Dauner-Lieb, '§ 93 AktG marginal number 24' in M Henssler and L Strohn (eds), *Gesellschaftsrecht* (5th ed. 2021); KJ Hopt and M Roth, '§ 93 marginal number 90' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); G Spindler, '§ 93 marginal number 69' in W Goette and M Habersack (eds), *'Münchener Kommentar zum AktG'* (5th ed. 2019); S Harbarth, 'Unternehmerisches Ermessen des Vorstands im Interessenkonflikt' in B Erle and others (eds), *'Festschrift für Peter Hommelhoff'* (2012) 323, 327; criticising this G Krieger and V Sailer-Coceani '§ 93 marginal number 19' in K Schmidt and M Lutter (eds), *AktG* (4th ed. 2020).

of AI applications can contribute to a reduction of transaction costs from an economic point of view and mitigate the principle-agent-conflict, as the interest of the firm will be aligned with decisions made by AI.[98] That is, AI can make the decision-making process (more) objective.[99] However, in order to achieve an actually objective result, the quality of the data used is decisive. If the data set itself is characterized by discriminatory or incorrect information, the result will also suffer from those weaknesses ('garbage in – garbage out'). Moreover, if the management board is in charge of developing AI applications inside the firm, it may have an interest in choosing experts and technology designs that favor its own benefit rather than the best interest of the company. This development could aggravate the principle-agent-conflict within the large public firm.[100]

IV. SUPERVISORY BOARD

For this reason, it will also be of fundamental importance in the future to have an institutional monitoring body in the form of the supervisory board, which enforces the interests of the company as an internal corporate governance system. With regard to the monitoring function, there is a distinction to be made as to whether the supervisory board makes use of AI itself while monitoring and advising the management of the company, or whether the supervisory board is monitoring and advising with regard to the use of AI by the management board.

1. *Use of AI by the Supervisory Board Itself*

As the members of the management board and of the supervisory board have to comply with the same basic standards of care and responsibility under sections 116(1) and 93(1)(1) AktG, the management board's AI related duties[101] essentially apply to the supervisory board accordingly. If the supervisory board is making an entrepreneurial decision, it can also rely on the business judgement rule.[102] This is true, for example, for the granting of approval for transactions requiring approval under section 111(4)(2) AktG, with regard to M&A transactions.[103] Furthermore, the supervisory board may use AI based personality and fitness checks when it appoints and dismisses management board members.[104] AI applications can help the supervisory board to structure the remuneration of the management board appropriately. They can also be useful for the supervisory board when auditing the accounting and in the compliance area, because they are able to analyze large amounts of data and uncover inconsistencies.[105]

2. *Monitoring of the Use of AI by the Management Board*

When it comes to the monitoring and advice on the use of AI by the management board, the supervisory board has to fulfil its general monitoring obligation under section 111(1) AktG. The starting point is the reporting from the management board under section 90 AktG.[106]

[98] Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1135.
[99] Cf. further Noack, 'Organisationspflichten' (n 10) 123.
[100] Cf. Enriques and Zetzsche, 'Corporate Technologies' (n 25) 42.
[101] See supra Section III 2.
[102] See supra Section III 3.
[103] Fleischer, 'Aufsichtsrat' (n 17) 121.
[104] For details, see I Erel and others, 'Selecting Directors Using Machine Learning' (*NBER*, March 2018) www.nber.org/papers/w24435.
[105] Cf. Fleischer, 'Algorithmen' (n 17) 121; Noack, 'Aufsichtsrat' (n 10) 507.
[106] Cf. Noack, 'Aufsichtsrat' (n 10) 502.

Namely strategic decisions on the leading guidelines of AI use is part of the intended business policy or at least another fundamental matter regarding the future conduct of the company's business according to section 90(1)(1) AktG. Furthermore, the usage of certain AI applications may be qualified as transactions that may have a material affect upon the profitability or liquidity of the company under section 90(1)(4) AktG. In this regard, the management board does not need to derive and trace the decision-making process of the AI in detail. Rather, it is sufficient for the management board to report to the supervisory board about the result found and how it specifically used the AI, monitored its functions, and checked the plausibility of the result.[107] In addition, pursuant to section 90(3) AktG, the supervisory board may require at any time a report from the management board on the affairs of the company, on the company's legal and business relationships with affiliated enterprises. This report may also deal with the AI related developments on the management board level and in other entities in a corporate group.

Finally, the supervisory board may inspect and examine the books and records of the company according to section 111(2)(1) AktG. It is undisputed that this also includes electronic recordings,[108] which the supervisory board can examine using AI in the form of a big data analysis.[109] Conversely, the supervisory board does not need to conduct its own inquiries using its information authority without sufficient cause or in the event of regular and orderly business development.[110] Contrary to what the literature suggests,[111] this applies even in the event that the supervisory body has unhindered access to the company's internal management information system.[112] The opposing view not only disregards the principle of a trusting cooperation between the management board and the supervisory board, but also surpasses the demands on the supervisory board members in terms of time.[113]

With a view to the monitoring standard, the supervisory board has to assess the management board's overall strategy as regards AI applications and especially systemic risks that result from the usage of AI in the company. This also comprises the monitoring of the AI-based management and organizational structure of the company.[114] If it recognizes violations of AI use by the management board, the supervisory board has to intervene using the general means of action. This may start with giving advice to the management board on how to optimize the AI strategy.

[107] Cf. Noack, 'Aufsichtsrat' (n 10) 502; for a different view, see Strohn, 'Rolle' (n 10) 374.

[108] S Hambloch-Gesinn and FJ Gesinn, '§ 111 marginal number 46' in W Hölters (ed), *AktG* (3rd ed. 2017); M Habersack, '§ 111 marginal number 74' in W Goette and M Habersack (eds), '*Münchener Kommentar zum AktG*' (5th ed. 2019); HC Grigoleit and L Tomasic, '§ 111 marginal number 49' in HC Grigoleit (ed), *AktG* (2nd ed. 2020).

[109] Noack, 'Organisationspflichten' (n 10) 140 *et seq*.

[110] HJ Mertens and A Cahn, '§ 111 marginal number 52' in W Zöllner and U Noack (eds), *Kölner Kommentar zum AktG* (3rd ed. 2010); A Cahn, 'Aufsichtsrat und Business Judgment Rule' (2013) *WM* 1293, 1299 (hereafter Cahn, 'Aufsichtsrat'); M Hoffmann-Becking, 'Das Recht des Aufsichtsrats zur Prüfung durch Sachverständige nach § 111 Abs 2 Satz 2 AktG' (2011) *ZGR* 136, 146 *et seq*.; M Winter, 'Die Verantwortlichkeit des Aufsichtsrats für "Corporate Compliance"' in P Kindler and others (eds), *Festschrift für Uwe Hüffer zum 70. Geburtstag* (2010) 1103, 1110 *et seq*.

[111] KJ Hopt and M Roth, '§ 111 marginal number 410' in H Hirte, PO Mülbert, and M Roth (eds), *Großkommentar zum AktG* (5th ed. 2015); W Zöllner, 'Aktienrechtliche Binnenkommunikation im Unternehmen' in U Noack and G Spindler (eds), *Unternehmensrecht und Internet* (2001) 69, 86.

[112] HJ Mertens and A Cahn, '§ 111 marginal number 52' in W Zöllner and U Noack (eds), *Kölner Kommentar zum* AktG (3rd ed. 2010); J Koch, '§ 111 marginal number 21' in U Hüffer and J Koch (eds), *Aktiengesetz* (14th ed. 2020); M Lutter, G Krieger, and D Verse, 'Rechte und Pflichten des Aufsichtsrats' (7th ed. 2020) marginal number 72; Cahn, 'Aufsichtsrat' (n 110) 1299; G Spindler, 'Von der Früherkennung von Risiken zum umfassenden Risikomanagement – zum Wandel des § 91 AktG unter europäischem Einfluss' in P Kindler and others (eds), *Festschrift für Uwe Hüffer zum 70. Geburtstag* (2010) 985, 997 *et seq*.

[113] For details, see Lieder, 'Entscheidungen' (n 76) 557 *et seq*., 560, 563.

[114] Cf. Wagner 'Legal Tech 2' (n 41) 1105; see furthermore Strohn, 'Rolle' (n 10) 375.

Furthermore, the supervisory board may establish an approval right with regard to the overall AI-based management structure. In addition, the supervisory board may draw personnel conclusions and install an AI expert on the management board level such as a CIO or CDO.[115]

V. CONCLUSION

AI is not the end of corporate governance as some authors predicted.[116] Rather, AI has the potential to change the overall corporate governance system significantly. As this chapter has shown, AI has the potential to improve corporate governance structures, especially when it comes to handling big data sets. At the same time, it poses challenges to the corporate management system, which must be met by carefully adapting the governance framework.[117] However, currently, there is no need for a strict AI regulation with a specific focus on corporations.[118] Rather, we see a creeping change from corporate governance to algorithm governance that has the potential to enhance, but also the risks to destabilize the current system. What we really need is the disclosure of information about a company's practices with regard to AI application, organization, and oversight as well as potentials and risks.[119] This kind of transparency would help to raise awareness and to enhance the overall algorithm governance system. For that purpose, the already mandatory corporate governance report that many jurisdictions require, such as the US,[120] the UK[121] and Germany,[122] should be supplemented with additional explanations on AI.[123]

In this report, the management board and the supervisory board should report on their overall strategy with regard to the use, organization, and monitoring of AI applications. This specifically relates to the responsibilities, competencies, and protective measures they established to prevent damage to the corporation. In addition, the boards should also be obliged to report on the ethical guidelines for a trustworthy use of AI.[124] In this regard, they may rely on the proposals drawn up on an international level. Of particular importance in this respect are the principles of the European Commission in its communication on 'Building Trust in Human-Centric Artificial Intelligence',[125] as well as the 'Principles on Artificial Intelligence' published by the OECD.[126] These principles require users to comply with organizational precautions in order to prevent

[115] See supra Section III 2(e).
[116] Leaning in that direction Armour and Eidenmüller, 'Kapitalgesellschaften' (n 9) 169 *et seq*.; cf. further, in general, V Boehme-Neßler, 'Die Macht der Algorithmen und die Ohnmacht des Rechts: Wie die Digitalisierung das Recht relativiert' (2017) *NJW* 3031 *et seq*.
[117] Cf. Enriques and Zetzsche, 'Corporate Technologies' (n 25) 42.
[118] More extensive Möslein, 'Digitalisierung' (n 17) 212; Weber, Kiefner, and Jobst, 'Unternehmensführung' (n 17) 1136; restrictive like here Armour and Eidenmüller, 'Kapitalgesellschaften' (n 9) 189; Enriques and Zetzsche, 'Corporate Technologies' (n 25) 47 *et seq*.; Noack, 'Organisationspflichten' (n 10) 142.
[119] Cf. Enriques and Zetzsche, 'Corporate Technologies' (n 25) 50 *et seq*.; Strohn, 'Rolle' (n 10) 377.
[120] NYSE, 'Listed Company Manual' Section 3, 303A.12 (*NYSE*, 25 November 2009) https://nyse.wolterskluwer.cloud/listed-company-manual.
[121] FCA, 'Listing Rules – FCA Handbook' LR 9.8.6.R (5) (*FCA*, January 2021) www.handbook.fca.org.uk/handbook/LR.pdf.
[122] *AktG*, section 161(1)(1).
[123] For inclusion in the code cf. also Noack, 'Organisationspflichten' (n 10) 113, 142.
[124] For precautionary compliance with the guidelines by the Supervisory Board, see Möslein, 'Digitalisierung im Aufsichtsrat: Überwachungsaufgaben bei Einsatz künstlicher Intelligenz' (2020) *Der Aufsichtsrat* 2(3).
[125] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee of the Regions, 'Building Trust in Human-Centric Artificial Intelligence'(*EUR-Lex*, 8 April 2019) https://eur-lex.europa.eu/legal-content/GA/TXT/?uri=CELEX%3A52019DC0168.
[126] OECD AI Policy Observatory, OECD Principles on Artificial Intelligence (*OECD.AI*, 2020) https://oecd.ai/en/ai-principles.

incorrect AI decisions, provide a minimum of technical proficiency, and ensure the preservation of human final decision-making authority. In addition, there is a safeguarding of individual rights, such as privacy, diversity, non-discrimination, fairness, and an orientation of AI to the common good, including sustainability, ecological responsibility, and overall societal and social impact. Even if these principles are not legally binding, a reporting obligation requires the management board and supervisory board to deal with the corresponding questions and to explain how they relate to them. It will make a difference and may lead to improvements if companies and their executives are aware of the importance of these principles in dealing with responsible AI.

20

Autonomization and Antitrust

On the Construal of the Cartel Prohibition in the Light of Algorithmic Collusion

Stefan Thomas

I. INTRODUCTION

The use of algorithms is associated with a risk of collusion. This bears on the construal of the cartel prohibition, on which the present chapter focuses. The hypothesis is that algorithms may achieve a collusive equilibrium without any involvement of natural persons. Against this backdrop, it is questionable whether and to what extent such an outcome can be qualified as a concerted practice in terms of the law.

The analysis will be structured as follows: first, it will be assessed in what way algorithms can influence competition on markets (Section II). Subsequently, the article will deal with the traditional criteria of distinction between explicit and tacit collusion, which might reveal a potential gap in the existing legal framework with respect to algorithmic collusion (Section III). Finally, it must be analyzed whether the cartel prohibition can be construed in a way that captures the phenomenon appropriately (Section IV). The chapter will close with a summary (Section V).

II. ALGORITHMIC COLLUSION AS A PHENOMENON ON MARKETS

It is widely accepted that the use of algorithms can precipitate collusive outcomes, at least in theory. There is no lack of attempts to systematize the different ways algorithms can be involved here. Since the first groundbreaking publications by *Ariel Ezrachi*, *Maurice Stucke*, and *Salil Mehra*, as well as other authors in the following, the matter has come into the focus of antitrust scholarship and practice.[1] Agencies have started to look

[1] See, e.g., A Ezrachi and M E Stucke, 'Sustainable and Unchallenged Algorithmic Tacit Collusion' (2020) 17 *Nw J Tech & Intell Prop* 217; A Ezrachi and ME Stucke, 'Algorithmic Collusion: Problems and Counter-Measures, Note to the OECD Roundtable on Algorithms and Collusion' (*OECD*, 31 May 2017) www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DAF/COMP/WD%282017%2925&docLanguage=En; A Ezrachi and ME Stucke, *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy* (2016); U Schwalbe, 'Algorithms, Machine Learning, and Collusion' (2018) 14 *J Comp L & Econ* 568; A Ittoo and N Petit, 'Algorithmic Pricing Agents and Tacit Collusion: A Technological Perspective' in H Jacquemin and A de Streel (eds), *L'intelligence artificiale et le droit* (2017) 241; SK Mehra, 'Antitrust and the Robo-Seller: Competition in the Time of Algorithms' (2016) 100 *Minn L Rev* 1323; VM Pereira, 'Algorithm-Driven Collusion: Pouring Old Wine Into New Bottles or New Wine Into Fresh Wineskins?' (2018) 39 *ECLR* 212; PG Picht and B Freund, 'Competition (Law) in the Era of Algorithms' (2018) 39 *ECLR* 403; VD Roman, 'Digital Markets and Pricing Algorithms – a Dynamic Approach towards Horizontal Competition' (2018) 39 *ECLR* 37; see for an assessment of the Commission's e-commerce sector inquiry regarding the risks of algorithmic collusion N Colombo, 'What the European Commission (Still) Does Not Tell Us about Pricing

into it.[2] First cases have emerged about restraints implemented on platforms involving computer technology. With respect to the United States (US) *Topkins*[3] must be mentioned, which involved alleged horizontal price fixing on a digital platform based on algorithms. For the European Union (EU), the *Eturas*[4] case comes to mind. The operator of a travel booking platform had informed the travel agencies using that platform that it intended to cap rebates granted to end-consumers. The European Court of Justice (ECJ) held that this amounted to a horizontally concerted practice among the travel agencies to the extent that these had not objected to this proposal. The Luxembourg competition authority in 2018 found the taxi booking app Webtaxi to be exempt from the cartel prohibition although the system involved an algorithmic horizontal alignment of prices. The agency found offsetting efficiencies to the benefit of consumers.[5] These cases have in common that the natural persons representing the companies involved were aware of the restrictions, or that they at least ought to have known about them. Algorithms were an element of implementing the restriction, yet the ultimate decision about the competitive restraint was taken by human beings. Under these conditions the cases did not pose severe difficulties in establishing explicit collusion. There is not a fundamental difference to cases in which parties communicate, for example, by way of hub-and-spoke cartelization through traditional means and forms of communication.[6]

A greater legal challenge is caused by the risk of autonomous algorithmic collusion. Computers with machine learning capabilities can possibly achieve or sustain a collusive equilibrium without any involvement of human knowledge or intent. The underlying scholarly discussion usually orbits around q-learning mechanisms.[7] The hypothesis is that algorithms with machine learning capabilities can act as computer agents exploring the success of their own actions, from which a collusive strategy can emerge as the optimum. In this event, it is

Algorithms in the Aftermath of the E-Commerce Sector Inquiry' (2018) 39 *ECLR* 478; See also P Pohlmann, 'Algorithmen als Kartellverstöße' in J Kokott and others (eds), *Europäisches, deutsches und internationales Kartellrecht, Festschrift für Dirk Schröder* (2018) 633, 645 *et seq*.; D Zimmer, 'Algorithmen, Kartellrecht und Regulierung' in J Kokott and others (eds), *Europäisches, deutsches und internationales Kartellrecht, Festschrift für Dirk Schröder* (2018) 999 *et seq*.

[2] *See* Autorité de la Concurrence and BKartA, 'Algorithms and Competition' (*BKartA*, November 2019) www.bundeskartellamt.de/SharedDocs/Publikation/EN/Berichte/Algorithms_and_Competition_Working-Paper.pdf?__blob=publicationFile&v=5; Autoridade da Concorrência, 'Paper on Digital Ecosystems, Big Data and Algorithms' (*AdC*, July 2019) www.concorrencia.pt/vEN/News_Events/Comunicados/Documents/Digital%20Ecosystems%20Executive%20Summary.pdf; Competition & Markets Authority, 'Pricing Algorithms: Economic Working Paper on the Use of Algorithms to Facilitate Collusion and Personalised Pricing' (*CMA*, October 2018) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/746353/Algorithms_econ_report.pdf.

[3] US Department of Justice, 'Former E-Commerce Executive Charged with Price Fixing in the Antitrust Division's First Online Marketplace Prosecution' (*US DoJ*, 6 April 2015) www.justice.gov/opa/pr/former-e-commerce-executive-charged-price-fixing-antitrust-divisions-first-online-marketplace.

[4] ECJ, Case C-74/14 *Eturas* (23 June 2016).

[5] Conseil de la Concurrence, 'Décision 2018-FO-01' (*Conseil de la Concurrence*, 7 June 2018) https://concurrence.public.lu/dam-assets/fr/decisions/ententes/2018/decision-n-2018-fo-01-du-7-juin-2018-version-non-confidentielle.pdf.

[6] In *Interstate Circuit v United States* from 1939 pricing restraints were implemented by vertical communication through analogue means, which, however, had the same effect as a digital communication would have had, *Interstate Circuit v United States* 306 US 208 (1939). On this case *see* BJ Rodger, 'The Oligopoly Problem and the Concept of Collective Dominance: EC Developments in the Light of U.S. Trends in Antitrust Law and Policy' (1995/1996) 2 *Colum J Eur L* 25, 30–36.

[7] This is a type of reinforcement-learning-algorithm, which adapts its conduct through experience. Learning takes place through the gaining of experience in these actions, which when proved successful, are repeated more frequently, while less successful actions are performed less frequently. Such a pattern allows the algorithms to develop a strategy that reaches the optimum or comes close to it. Therefore, q-learning allows an optimization without prior knowledge of the problem which is to be solved.

conceivable that even the programmer of the algorithm was not aware of this potential outcome.[8] The market effect, therefore, can be a collusive equilibrium, albeit absent any human involvement.

Lawyers, economists, and computer scientists are still at odds over the likeliness and actual occurrence of autonomous algorithmic collusion. Some consider it a realistic scenario that tends to be underestimated.[9] The German Bundeskartellamt and the French Autorité de la Concurrence have refrained from a definitive conclusion so far.[10] *Ulrich Schwalbe*, in his seminal article, points out that the game theoretical dilemma that has to be solved by autonomous computer agents to achieve a stable collusive equilibrium is huge and cannot be easily overcome in practice.[11] A more recent study by *Emilio Calvano* and others[12], however, concludes that q-learning algorithms, in fact, can autonomously collude. The EU Commission, in its proposal for a 'New Competition Tool', mentions the risk that digital platforms can create ecosystems in which collusion arises, which can be read as a recognition of the phenomenon as a matter of concern.[13] Against this backdrop, it is expedient to elaborate further on the application of the cartel prohibition in such cases.

III. ON THE SCOPE OF THE CARTEL PROHIBITION AND ITS TRADITIONAL CONSTRUAL

The conceptual problem behind the traditional construal of the cartel provision is the difference in the structure of the law on the one hand and the economic determinants for collusive equilibria on the other.[14] Anticompetitive collusive equilibria are characterized by the fact that the participants consider it individually rational to pursue such a strategy. That is the case for types of collusion that are usually referred to as explicit and which are illegal in the same way as it holds true for so-called tacit collusion, which is seen as not to fall within the scope of the prohibition. Agreements in breach of the cartel prohibition are null and void so that they cannot

[8] S Thomas, 'Harmful Signals: Cartel Prohibition and Oligopoly Theory in the Age of Machine Learning' (2019) 15 *J Comp L & Econ* 159.

[9] A Ezrachi and ME Stucke, 'Sustainable and Unchallenged Algorithmic Tacit Collusion' (2020) 17 *Nw J Tech & Intell Prop* 217.

[10] Autorité de la Concurrence/BKartA, 'Algorithms and Competition, Nov 2019' (*BKartA*, November 2019) www.bundeskartellamt.de/SharedDocs/Publikation/EN/Berichte/Algorithms_and_Competition_Working-Paper.pdf?__blob=publicationFile&v=5; Autoridade da Concorrência, 'Paper on Digital Ecosystems, Big Data and Algorithms, July 2019, Executive Summary' (*AdC*, July 2019) www.concorrencia.pt/vEN/News_Events/Comunicados/Documents/Digital%20Ecosystems%20Executive%20Summary.pdf.

[11] U Schwalbe, 'Algorithms, Machine Learning, and Collusion' (2018) 14 *J Comp L & Econ* 568.

[12] E Calvano and others, 'Artificial Intelligence, Algorithmic Pricing, and Collusion' (2020) 110 *Am Econ Rev* 3267. The authors present a study on the capability of q-learning algorithms to achieve equilibria. They come to the conclusion that algorithms can learn to implement anticompetitive pricing.

[13] See the explanation given in the Inception Impact Assessment for a 'New Competition Tool', S. 1: 'The Commission's enforcement experience in both antitrust and merger cases in various industries points to the existence of structural competition problems that cannot be tackled under the EU competition rules while resulting in inefficient market outcomes. [...] Even short of individual market power, increasingly concentrated markets can allow companies to monitor the behaviour of their competitors and create incentives to compete less vigorously without any direct coordination (so-called tacit collusion). Moreover, the growing availability of algorithm-based technological solutions, which facilitate the monitoring of competitors' conduct and create increased market transparency, may result in the same risk even in less concentrated markets.' *see* European Commission, 'Inception Impact Assessment' (*EC*, 29 May 2020) https://ec.europa.eu/competition/consultations/2020_new_comp_tool/new_comp_tool_inception_impact_assessment.pdf.

[14] The following refers to Article 101 TFEU, yet the same problems arise under the cartel provisions of many other jurisdictions in a similar way.

be enforced. Therefore, any cartel is only stable so long as the firms participating in it consider it rational to remain involved. All types of collusive equilibria can, therefore, be considered as non-cooperative games in terms of game theory.[15] Yet still, the law does not prohibit the achievement or sustaining of a collusive equilibrium as such. Instead, the provision is confined to certain types of measures, which are described as agreement, concerted practice, or decision. Mere tacit collusion is supposed to be distinct from these aforementioned types of anticompetitive conduct. While explicit collusion (i.e. achieved by agreement), concerted practice, or decision, is prohibited, tacit collusion is found to be legitimate.

As becomes obvious, the traditional construal of the law rests on a description of the means and forms by which firms interact when defining collusion. In the context at hand, the category of concerted practices has the greatest relevance. It is conceived of as something whereby a 'practical cooperation' is substituted for the risks of competition. Such practical cooperation, in turn, is supposed to be different from merely observing a rival's conduct and reacting to it.[16] Similar approaches of distinction apply under section 1 of the US Sherman Act.[17] There, so-called conscious parallelism is deemed not to fall within the scope of the prohibition.[18] For the finding of a cartel, so-called plus factors[19], or 'facilitating practices'/'facilitating devices' need to be established.[20]

It is argued that, whereas firms when tacitly colluding merely observe each other and react independently, the mechanism allegedly differs if they opt for a practical cooperation. A private exchange of pricing information can serve as an example for the latter. The jurisprudence of the courts requires that such practical cooperation must be substituted 'knowingly' for the risks of competition for it to amount to a concerted practice.[21] The concept, therefore, hinges on the inner sphere of the firms involved.

Against the afore, it is questionable whether autonomous algorithmic collusion is prohibited under the traditional enforcement paradigms. If the firms lack of knowledge or intent with

[15] See, e.g., J Friedmann, *Game Theory with Applications to Economics* (1986) 184: 'The fundamental distinction between cooperative and noncooperative games is that cooperative games allow binding agreements while noncooperative games do not.'; L Kaplow, *Competition Policy and Price Fixing* (2013) 177; see also E J Green and R H Porter, 'Noncooperative Collusion under Imperfect Price Information' (1984) 52 *Econometrica* 87; D G Baird and others, *Game Theory and the Law* (1994) 165–178.

[16] ECJ, Case 48-69 *Imperial Chemical Industries Ltd. v Commission of the European Commission* [1972] paras 64 and 65; see also ECJ, joined cases C-89/85 and others *A. Ahlström Osakeyhtiö and Others v European Communities* [1993] para 63; ECJ, case C-8/08 *T-Mobile Netherlands BV and others v Raad van bestuur van de Nederlandse Mededingingsautorieteit* [2009] para 26.

[17] Act of July 2, 1890 (Sherman Anti-Trust Act) 15 U.S. Code § 1.

[18] *Theatre Enterprises v Paramount* 346 US 537 (1945); RH Bork, *The Antitrust Paradox* (1978) 178 *et seq*.; also arguing in favor of a distinction between 'illegal agreement' and 'conscious parallelism' MD Blechman, 'Conscious Parallelism, Signaling and Facilitating Devices: The Problem of Tacit Collusion under the Antitrust Laws' (1979) 24 *NYL Sch L Rev* 881, 882, 889.

[19] The notion 'plus factor' was, reportedly, used for the first time in this context in *C-O-Two Fire Equip. Co. v. United States* 197 F2d 489, 493 (9th Cir.), cert. denied, 344 U.S. 892 (1952); see on that MD Blechman, 'Conscious Parallelism, Signaling and Facilitating Devices: The Problem of Tacit Collusion under the Antitrust Laws' (1979) 24 *NYL Sch L Rev* 881, 885.

[20] The US DoJ has defined 'facilitating devices' as 'mechanisms that facilitate the achievement of an industry pricing or output consensus and police deviations from it [in concentrated industries].' See US DoJ, 'Memorandum of John H Shenefield, Assistant Attorney General, Antitrust Division, Shared Monopolies' (1978) 874 *Antitrust & Trade Reg Rep (BNA)* at F-1. See also GA Hay, 'Facilitating Practices: The Ethyl Case (1984)' in JE Kwoka and LJ White (eds), *The Antitrust Revolution: Economics, Competition, and Policy* (3rd ed. 1999) 182–201.

[21] ECJ, Case 48-69 *Imperial Chemical Industries Ltd. v Commission of the European Commission* (14 July 1972) paras 64 and 65; see also ECJ, Joined Cases C-89/85 and others *Ahlström Osakeyhtiö and Others v European Communities* (20 January 1994) para 63; ECJ, Case C-8/08 *T-Mobile Netherlands BV and others v Raad van bestuur van de Nederlandse Mededingingsautorieteit* [2009] para 26.

respect to the fact that their computer agents pursue a collusive strategy, it cannot be said that these firms 'knowingly' substitute a practical cooperation for competition. Several authors, therefore, point to the risk that the cartel prohibition might stop short of preventing such outcomes. *Ezrachi and Stucke* argue that collusion achieved by machine learning systems can fall outside the scope of the cartel prohibition for 'lack of evidence of an anticompetitive agreement or intent.'[22] In a similar vein, *Calvano* and others conclude from their study[23]:

> From the standpoint of competition policy, these findings should probably ring an alarm bell. Today, the prevalent approach to tacit collusion is relatively lenient, in part because tacit collusion among human decision-makers is regarded as extremely difficult to achieve. While we have no direct comparative evidence for algorithms relative to humans, our results suggest that algorithmic collusion might not be that improbable. If this is so, then the advent of algorithmic pricing could well heighten the risk that tolerant antitrust policy will produce too many false negatives.

Some authors, therefore, highlight that the enforcement paradigms might warrant amendments to close such regulatory gaps.[24]

IV. APPROACHES FOR CLOSING LEGAL GAPS

1. *On the Idea of Personifying Algorithms*

It is questionable whether the potential enforcement gap in antitrust[25] can be overcome by defining algorithms as 'undertakings'. The notion of an undertaking in EU antitrust, indeed, is a very broad concept which has many facets and functions. As a working definition that applies to the most common types, an undertaking can be described as a combination of assets and people that act on a market governed by a management body or further representatives irrespective of the legal personality or corporate structure. For an undertaking to act, or to have knowledge or intent, it is the action, the knowledge, or the intent of the human beings representing it which is attributed to it. If a company manager knowingly enters into an exchange of sensitive pricing information with the manager of a rival, it can, therefore, be said that these 'undertakings' substituted a practical cooperation for the risks of competition.

Yet what substantive meaning would terminology such as a 'practical cooperation' or 'knowingly' have, if they were applied to an algorithm? The problem is that the legal terminology is coined on human interaction and cognition, so that it will be vastly deprived of its meaning if transferred to a computer system. How shall an action of an algorithm be identified as 'knowingly', as opposed to another action that is supposed to happen un-knowingly? In what way is it meaningful to consider the actions of an algorithm as a 'practical cooperation', as opposed to a mere intelligent adaption to information obtained by this algorithm on the market? To rely on such human concepts of cognition with respect to the regulation of algorithms will likely end up in semantic exercises with limited substance.

[22] A Ezrachi and M E Stucke, 'Artificial Intelligence & Collusion: When Computers Inhibit Competition' (2017) 1775, 1796 *U Ill L Rev*.

[23] E Calvano and others, 'Artificial Intelligence, Algorithmic Pricing, and Collusion' (2020) 110 *Am Econ Rev* 3267, 3295.

[24] C Veljanovski, *Cartel Damages: Principles, Measurement & Economics* (2020) 100 para 7.07: 'the law may need to be applied in a different fashion.'

[25] On the attribution of legal personality to algorithms as a general legal issue see H Eidenmüller, 'The Rise of Robots and the Law of Humans' (2017) 4 *ZEuP* 765.

2. *On the Idea of a Prohibition of Tacit Collusion*

Another way of dealing with the problem could be in equating tacit collusion and explicit collusion.[26] This would mean that under antitrust law any collusive strategy would qualify as an illicit cooperation so that any further distinctions based on the inner sphere of the persons involved would become obsolete. Such view has been suggested in the past independently of the issue of algorithmic collusion. Notable proponents were *Richard Posner* in his earlier writings[27] (he has since changed his view[28]), *Richard Markovits*[29], and more recently *Louis Kaplow*.[30]

One might feel inclined to hold that such a view does not reconcile with the structure of the law. Yet it is questionable whether this counterargument would be very strong. The notion of concerted practices can possibly be construed in a way as to extend to cases in which the collusive outcome is based on a mechanism of observation and retaliation, which is characteristic for tacit collusion as it is for explicit collusion. As outlined earlier, both categories share the feature that they can be described as a non-cooperative game in terms of game theory. It is, therefore, rather a semantic issue whether some ways of engaging in such a non-cooperative strategy can be tagged as a 'practical cooperation' or not, while the underlying economic principles remain the same. Firms observe and react in ways that are deemed individually optimal irrespective of which words are used to describe this phenomenon. Especially in the grey area between typical cases of explicit cooperation on the one hand and tacit oligopoly conduct on the other, it becomes apparent how brittle the traditional concept of distinction is.[31] Consider that even in cases that would usually be qualified as tacit collusion, firms cooperate in that they observe each other and react to the information they have obtained from observing

[26] For a critical review of this view see P Pohlmann, 'Algorithmen als Kartellverstöße' in J Kokott and others (eds), *Europäisches, deutsches und internationales Kartellrecht, Festschrift für Dirk Schroeder* (2018) 633, 645 *et seq*.

[27] RA Posner, 'Oligopoly and the Antitrust Laws: A Suggested Approach' (1969) 21 *Stan L Rev* 1562, 1575: 'the tacit colluder should be punished like the express colluder.'; RA Posner, 'Oligopolistic Pricing Suits, the Sherman Act, and Economic Welfare'(1976) 28 *Stan L Rev* 903.

[28] Posner has meanwhile distanced himself from this view and takes the opposite position according to which tacit collusion should not be equated with explicit collusion, see RA Posner, 'Review of Kaplow, Competition Policy and Price Fixing' (2014) 79 *Antitrust LJ* 761; on that see also CS Hemphill, 'Posner on Vertical Restraints' (2019) 86 *U Chi L Rev* 1057, 1073.

[29] Markovits wants to distinguish between 'normal' or 'natural' oligopolistic pricing on the one hand and 'contrived' oligopolistic pricing on the other, see RS Markovits, 'A Response to Professor Posner' (1976) 28 *Stan L Rev* 919, 933–934; RS Markovits, 'Oligopolistic Pricing Suits, the Sherman Act, and Economic Welfare, Part II: Injurious Oligopolistic Pricing Sequences: Their Description, Interpretation, and Legality under the Sherman Act'(1974) 26 *Stan L Rev* 717, 738; see also RS Markovits, 'Oligopolistic Pricing Suits, the Sherman Act, and Economic Welfare, Part III: Proving (Illegal) Oligopolistic Pricing: A Description of the Necessary Evidence and a Critique of the Received Wisdom about Its Character and Cost'(1975) 27 *Stan L Rev* 307, 315–319; RS Markovits, 'Oligopolistic Pricing Suits, the Sherman Act, and Economic Welfare, Part IV: The Allocative Efficiency and Overall Desirability of Oligopolistic Pricing Suits'(1975) 28 *Stan L Rev* 45, 44–60. Posner criticizes this distinction as suggested by Markovits, see RA Posner, 'Oligopolistic Pricing Suits, the Sherman Act, and Economic Welfare, A Reply to Professor Markovits' (1976) 28 *Stan L Rev* 903, 908 and 913 *et seq*.

[30] L Kaplow, 'An Economic Approach to Price Fixing' (2011) 77 *Antitrust LJ* 343, 350; see also L Kaplow, 'Direct Versus Communications-Based Prohibitions on Price Fixing' (2011) 3 *J Legal Analysis* 449; L Kaplow, 'On the Meaning of Horizontal Agreements in Competition Law'(2011) 99 *Calif L Rev* 683; L Kaplow, *Competition Policy and Price Fixing* (2013). On this strand of arguments see also D Zimmer, 'Kartellrecht und neuere Erkenntnisse der Spieltheorie: Vorzüge und Nachteile einer alternativen Interpretation des Verbots abgestimmten Verhaltens (§ 25 Abs 1 GWB, Art 85 Abs 1 EWGV)' (1990) 154 *ZHR* 470.

[31] On that see also Nicolas Petit's suggestion of remedies against tacit collusion and the idea of applying a form of equivalence with express collusion: N Petit, 'Re-Pricing through Disruption in Oligopolies with Tacit Collusion: A Framework for Abuse of Collective Dominance' (2016) 119–138 World Competition; N Petit, 'The Oligopoly Problem in EU Competition Law' in I Liannos and D Geradin (eds), *Research Handbook in European Competition Law* (2013) 259–349.

each other. One might ask the question: In what way is this not a 'practical cooperation'? Also, a collusive equilibrium can bear negatively on consumer welfare[32] irrespective of the means and forms used by firms to sustain it. This seems to add to the argument that the distinction between tacit and explicit collusion is of limited expedience.

Yet still the idea of equating tacit collusion with explicit collusion faces severe objections, which might explain why it has not gained more recognition among scholars and enforcers.[33]

Any law must provide the addressee the opportunity to abide by it through choosing a compliant course of action. Otherwise, the law would be perplexing. Recall, now, that collusion in the realm of antitrust is a non-cooperative game.[34] This means that the strategy is individually rational for each participant. Sanctioning collusion without more, therefore, would amount to prohibiting the pursuit of an individually rational strategy. *Kaplow* reacts to this objection by pointing out that it is a common feature of the legal order to prohibit types of conduct that are individually rational[35], such as the stealing of an apple.[36] By imposing a sanction, the law ensures that it becomes rational to refrain from that course of action, he advances.

This argument, however, is not able to overcome the conceptual problems that would arise if the law prohibited the collusive outcome as such. While it is perfectly clear what a person must do to 'not steal an apple', it is much less obvious what course of action a firm would have to take in order to 'not pursue a collusive strategy'. The negatory of stealing an apple is unambiguous. The course of action in order to avoid the sanction is to not steal the apple. With a prohibition of a collusive market outcome, however, it would be much more difficult to describe, in an unambiguous way, what the compliant course of action would be. One of the reasons for this is that, from an *ex ante* perspective, it is unclear what outcome a collusive strategy among firms might produce. If that is not known, however, it is equally unclear what price a firm must set in order to not charge at a collusive level. If the firms do not know whether a collusive strategy would yield a market price of € 5 or rather of merely € 4, they cannot know *ex ante* whether individually charging a price of € 4 would be 'non-collusive' or not.

On a more philosophical level, another objection buoys. Effectively, the law would require the addressee to pursue any market strategy so long as it is not rational, assumed that the rational strategy would be collusion. It is questionable, however, whether an addressee of the law can be required to intentionally act irrationally. Any legal prohibition must conceive of the addressee as an intelligent entity, for if the addressee were not intelligent, it could not abide by the law in the first place. From a philosophical perspective, it is unclear, however, whether a person can 'rationally act irrationally'. Can a firm be obliged to randomize prices in order to protect itself from being accused of acting rationally-collusive?

Yet even ignoring this fundamental problem and moreover assuming it were possible to define a hypothetical collusive price level *ex ante*, it would remain unclear whether a non-collusive strategy could be defined with a sufficient degree of precision so that firms could abide by the law. Would it suffice to undercut the hypothetical collusive price by, for example, 2%? Or

[32] Or other competitive parameters, such as quality.

[33] S Thomas, 'Harmful Signals: Cartel Prohibition and Oligopoly Theory in the Age of Machine Learning' (2019) 15 *J Comp L & Econ* 159; S Thomas, 'Herausforderungen des Plattformwettbewerbs für das Kartellverbot' in S Thomas and others (eds), *Das Unternehmen in der Wettbewerbsordnung, Festschrift für Gerhard Wiedemann zum 70. Geburtstag* (2020) 99 *et seq*; S Thomas, 'Horizontal Restraints on Platforms: How Digital Ecosystems Nudge into Rethinking the Construal of the Cartel Prohibition' (2021) 44 *World Competition* 53. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645095.

[34] See supra n 15.

[35] Setting aside religious beliefs and ethical convictions of the individual.

[36] L Kaplow, 'An Economic Approach to Price Fixing' (2011) 77 *Antitrust LJ* 343, 431.

would the law require every firm to price at marginal cost, for under perfect competition this would be the hypothetical competitive price, even though perfect competition usually does not exist? Merely imposing a sanction on the pursuit of a collusive strategy does not solve any of these conceptual problems.[37]

3. *Harmful Informational Signals As Point of Reference for Cartel Conduct*

a. Conceptualization

Another conceptual venture to solve the issue could be to maintain a distinction between illicit cartel conduct and legitimate coordination, yet to substitute a new criterion for the traditional paradigm, viz. for the practical cooperation adage. Such alternative criterion of distinction could be checking the informational signals released by colluding firms for their propensity to create a net consumer harm. Accordingly, an illicit concerted practice would be found if firms, or the algorithms relied on by firms, released informational signals which reduced consumer rent if compared to a counterfactual in which such informational signals were absent. The counterfactual, therefore, would not be a hypothetical market without collusion. It would be the same market absent a particular informational signal.[38] To clarify the specifics of this approach, and to contrast it with the view expressed by *Kaplow* and others, the following explanations shall be made.

In contrast to a situation where any collusive equilibrium is prohibited, so that it is unclear which course of action to take in order to abide by the law, confining the prohibition to harmful signals leaves the addressee a binary choice that is conceptually clear: an informational signal that is ultimately harmful to consumers can either be released or not. There is no element of imposed irrationality in such a prohibition. To refrain from the release of a harmful informational signal can also mean that the addressee must choose to release a different signal in order to avoid a harmful effect. To exemplify the idea, reference can be made to the EU Commission's remedy decision in *Container Shipping*, where public price announcements were not abandoned completely but limited in scope in order to avoid the creation of a consumer harm.[39]

A relevant signal in this sense can be any release of market-related information independent of whether it takes place publicly or in private, whether consumers are involved or not. Even price lists, price announcements, or other types of public display of prices or other parameters can qualify as a signal that ultimately leads to a collusive equilibrium. In contrast to the aforementioned opinion of *Kaplow*, however, this would not suffice for the conclusion that a restriction of competition in terms of the cartel prohibition is established. Rather, the release of such informational signals would fall outside the scope of Article 101(1) TFEU to the extent that it produces offsetting consumer benefits.

This reflects the fact that there is a plethora of cases where public market information leads to collusive equilibria, although the benefit to consumers being derived from this information outweighs the gross harm of the collusive outcome if compared to a situation in which such informational signals were not made. The reason why, in most cases, public price lists and similar forms of information are not prohibited despite their risk to precipitate collusive

[37] Also voicing concerns E Elhauge and D Geradin, *Global Competition Law & Economics Chapter 6 C* (2nd ed. 2011) 843.

[38] S Thomas, 'Harmful Signals: Cartel Prohibition and Oligopoly Theory in the Age of Machine Learning' (2019) 15 *J Comp L & Econ* 159.

[39] EU Commission, decision of 7 July 2016 *Container Shipping* AT 39850.

equilibria lies in the fact that they usually bring about benefits to consumers that offset the potential harm.[40] If firms refrain from the release of such pricing signals, the distribution of goods might be significantly impeded. Consumers can face difficulties in planning their purchases ahead or in pursuing multisource strategies. It can be said, therefore, that sometimes 'vertical transparency' creates a greater benefit to consumers than what is taken away by the horizontal collusion that concomitantly results from it. It is necessary, therefore, to balance both effects in order to get a full picture of the economic impact of an informational signal. If such an analysis of the economic effects is integrated into the assessment of a concerted practice, this creates a system in which harm and benefit of a measure can be distinguished without the need to make recourse to notions such as the 'practical cooperation' adage or the intention of the firms involved.

This concept would allow an entity to deal with the phenomenon of autonomous algorithmic collusion under the cartel prohibition. Achieving or sustaining such a market outcome would be prohibited if and to the extent that the harm to consumers were greater than the benefits associated with the release of the underlying informational signals.[41] If algorithms achieved a collusive equilibrium by communicating with each other and without this interaction providing any useful information to consumers, this could, therefore, constitute a concerted practice in terms of the law. If, on the other hand, a digital platform aggregated and provided information to consumers in a way that benefits them, and if the efficiency potential of this platform could not materialize without this release of information, Article 101(1) TFEU would not be triggered even though the conduct might, ultimately, give rise to a collusive equilibrium, if and to the extent that the benefit of the former offsets the harm of the latter. The counterfactual, therefore, would not be a digital platform without collusion. It would be a situation in which the informational signals were not released. If the platform, in such a counterfactual scenario, were not operated or operated with a lesser efficiency, consumers could be deprived of the benefits resulting from it. This could mean a smaller range of suppliers being visible to them, less information being available to help consumers plan their purchases, etc.

This demonstrates that a collusive outcome can be an inevitable consequence of an algorithm-based system that produces benefits, although horizontal restraints, even on prices, are involved. The decision of the Luxembourg competition agency has made this clear with respect to the taxi app Webtaxi.[42] The agency found consumer benefits in the fact that the platform improved on the supplies with transportation services, even though a horizontal collusion on prices was an inevitable side effect. While the decision accounted for these efficiencies within the scope of an exemption from the cartel prohibition, the concept presented here would integrate this analysis already into the assessment of a concerted practice. As previously outlined, this is a consequence resulting from the substitution of an effects analysis for the less useful 'practical cooperation adage' in order to discriminate between legitimate and illicit collusion in algorithm cases.

[40] OECD, 'Background Paper, Policy Roundtables on Unilateral Disclosure of Information with Anticompetitive Effects' (*OECD*, 11 October 2012) paras 1 and 2.3.1. www.oecd.org/daf/competition/Unilateraldisclosureofinformation2012.pdf.

[41] S Thomas, 'Harmful Signals: Cartel Prohibition and Oligopoly Theory in the Age of Machine Learning' (2019) 15 *J Comp L & Econ* 159.

[42] See supra n 5.

b. Possible Objections

Such a concept of harmful informational signals, of course, provokes objections. They shall be dealt with in the remainder of this chapter. One might want to invoke that this construal of the law does not reconcile with the structure of the provision as shaped by the jurisprudence. Admittedly, the courts have not yet recognized an interpretation as suggested here. Rather, the Court of Justice, currently, relies on the notion of 'practical cooperation' in order to distinguish between concerted practices and tacit collusion. The established set of criteria does not contain a place for an economic effects assessment. On the other hand, the Court of Justice has not yet had an opportunity to hone the law with respect to the phenomenon of autonomous algorithmic collusion. It is, therefore, conceivable that the courts, upon preparatory enforcement steps done by the agencies, ultimately consider the option to readjust some of the enforcement paradigms in order to close potential regulatory gaps.

Also, it should be noted that even under the current decisional practice an effects analysis can be part of the assessment of Article 101(1) TFEU. As to the distinction between restrictions by effect and those by object, the potential of the measure to produce restrictive effects bears significance.[43] In that regard, the Court of Justice has made clear that, among other things, it must be analyzed whether and in what way consumers might suffer from the measure at stake.[44] This line of reasoning already mirrors an effects analysis based on the consumer welfare paradigm, which also is the backbone of the present proposal. In a similar way, the Commission practice demonstrates that the notion of a competitive restraint involves an analysis of the effects on consumer rent. Even though the Court of Justice has ruled that, for a restriction of competition to arise, it is unnecessary to demonstrate concrete consumer harm[45], the Commission takes this into account if it helps to discriminate between legitimate and illicit conduct. The Commission, for example, argues that even horizontal price fixing, depending on the structure and function of the cooperation, might warrant a close examination within the 'by effect category', which demonstrates that it is possible to conceive of cases where such conduct falls outside of the scope of Article 101(1) TFEU without any further examination of the legal exemption rule in Article 101(3) TFEU.[46]

On a practical level, one might want to invoke that the assessment of whether an informational signal precipitates a net consumer harm or not is too complicated to rely on as an enforcement paradigm. Yet it must be noted that even the currently existing decisional practice is not void of elements of an effects analysis, as demonstrated previously. Beyond that, firms and enforcers face exactly these difficulties already within the realm of Article 101(3) TFEU, so that it could not be said that such difficulties are idiosyncratic to the proposal made here.

Beyond that, one might raise the question in what way firms can become responsible for the conduct of an algorithm, if the strategy pursued by the latter is unknown to the former. In legal terms, however, it is possible to hold someone responsible for the organization of an enterprise. Firms, therefore, could be considered obliged to terminate the use of an algorithm or to alter its

[43] ECJ, Case C-32/11 *Allianz Hungária Biztosíto v Gazdásagi Versenyhivatal* (14 March 2013) para 66.

[44] ECJ, Case C-67/13 P *Groupement des Cartes Bancaires v European Commission* (11 September 2014) para 51.

[45] ECJ, Joined Cases C-501/06 P and others *GlaxoSmithKline Services Unlimited v European Commission* (6 October 2009) para 63.

[46] See, e.g., EU Commission, 'Guidelines on the Applicability of Article 101 of the Treaty on the Functioning of the European Union to Horizontal Co-Operation Agreements' (2011) OJ 2011 C 11/1 para 161, where the Commission explains that in the case of a joint distribution which is downstream to a joint production, horizontal price fixing can be assessed within the 'by effect category'. While this situation is not identical with the case of algorithmic collusion, it demonstrates that it is not conceptually impossible to turn towards the effects on consumer rent when evaluating whether a certain conduct amounts to a breach of Article 101(1) TFEU or not.

paradigms, if and to the extent that it produces more harm than benefit to consumers. Firms could, therefore, be ordered, by way of an administrative decision, to make such adjustments to their business strategy either by tweaking the algorithm or by stopping its use altogether. Antitrust economists already expound ways to design platforms in a way that counters collusive risks emerging on it from machine learning systems. *Justin Johnson*, *Andrew Rhodes*, and *Matthijs Wildenbeest* published a study in 2020 on how a choice algorithm on a sales platform can impact the likeliness of collusion among independently acting q-learning algorithms.[47] It goes without saying that the imposition of a fine, or a liability for damages would, in any event, require negligence or intent on the firm's part. That, in turn, would require the agency to establish that some degree of knowledge or intent with respect to the achievement of a collusive equilibrium can be established among the actual firms that relied on the algorithm. Such would be absent if the equilibrium were achieved or sustained independently by a machine learning process not guided or anticipated by the firms that rely on the outcome.[48] Yet still, in such an event a mere administrative order, absent any sanction or damages award, could be issued.

Finally, one might want to invoke that it would be disproportionate to make such far-reaching amendments to the construal of the cartel prohibition for the sole purpose of closing a regulatory lacuna with respect to algorithms. It is not the intention behind this proposal, however, to render the established enforcement paradigms obsolete in their entirety. Rather, the suggestion made here should be conceived of as a mere addition to the established principles that would still bear relevance in the majority of non-algorithmic collusion cases. A private exchange of pricing information between the managers of two rivals, for example, could still be considered a practical cooperation without further effects analysis required, for its potential to precipitate a consumer harm, as reflected in the Commission's Horizontal Guidelines.[49] The informational signal approach suggested here, on the other hand, could become relevant if, in an algorithm case, the traditional criteria did not allow the application of the law in a meaningful way. The present suggestion, therefore, is meant as a humble contribution to existing paradigms, not as a postulate of a total substitution for them.

V. CONCLUSION

This article is a conceptual sketch of a way to deal with the intricacies coming along with autonomous algorithmic collusion. Such risk is being discussed, especially, with respect to q-learning algorithms. Even though practical cases have not yet emerged, there is sufficient reason to address potential issues as a precautionary measure at this point in the scholarly debate. It was the intention to demonstrate that the traditional construal of the law, which relies on a description of human behavior, appears inapt for effectively tackling machine-induced equilibria. Applying the established criteria in such cases would very likely lead to a harping on words without substance. There is simply no point in venturing to assess whether algorithms

[47] J Johnson and others, 'Platform Design when Sellers Use Pricing Algorithms' (*SSRN*, 12 September 2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3691621.

[48] The problem of how to detect algorithmic collusion is independent of the concept of the notion of a concerted practice. Enforcers will struggle with a total absence of direct evidence if the algorithmic computer agents acted independently of human interaction. Market comparison methods, however, could be used to find out about the competitiveness of the pricing level. This is an area with a great demand for further research. Yet the problems to detect algorithmic collusion do not call into question the need to expound ways how the law should be construed in the event that algorithmic collusion can be found.

[49] EU Commission, 'Guidelines on the Applicability of Article 101 of the Treaty on the Functioning of the European Union to Horizontal Co-Operation Agreements' (2011) OJ 2011 C 11/1 para 94.

'practically cooperated' as opposed to 'merely observed each other'. The present proposal, therefore, is intended to serve as a conceptualization of the notion of concerted practices for cases that would otherwise elude the cartel prohibition. To conclude, the entire problem of autonomous algorithmic collusion is an example of the necessity for interdisciplinary research between lawyers, economists, and computer scientists. At the same time, the problem highlights how enforcement paradigms, that hinge on descriptions of the inner sphere and conduct of human beings, may collapse when applied to the effects precipitated by independent computer agents. The subject matter of this chapter is, therefore, an example for the greater challenges that the entire legal order faces in light of the progress of machine learning.

21

Artificial Intelligence in Financial Services

New Risks and the Need for More Regulation?

*Matthias Paul**

I. INTRODUCTION

The financial services industry has been at the forefront of digitization and big data usage for decades. For the most part, data processing has been automized by information management systems. Not surprisingly, Artificial Intelligence (AI) applications, capturing the more intelligent ways of handling financial activities and information, have increasingly found their way into the financial services industry over the last years; from algorithmic trading, smart automized credit decisions, intelligent credit card fraud detection processes, personalized banking applications and even into areas like so-called robo-advisory services and quantitative investment and asset management more recently.[1]

The financial industry has also been one of the most regulated industries in the world. In particular, since the collapse of Lehmann Brothers in 2008, leading into one of the most severe financial crises in history, regulation efforts of all kinds of finance-related activities and financial organizations as a whole by the different regulators around the world have significantly increased. In general, most regulations relating to the financial industry, in particular those put into place after the financial crisis in 2008, have focused on areas like safeguarding the financial institutions themselves, safeguarding the customers of financial institutions, and making sure the institutions comply with general laws overall and on a global scale, given the truly global nature of the financial industry.

More recently, authors have argued that with the emergence of AI-based applications in the financial industry, new kinds of risks have emerged that require additional regulations.[2] They have pointed for instance to increased data processing risk, cybersecurity risks, additional challenges to financial stability, and even to general ethical risks stemming from AI in financial services. Some regulators like the Monetary Authority of Singapore (MAS) have proposed an AI

* I want to thank Silja Voeneky for many insightful discussions of the topic of AI, for sharing and exchanging many ideas, and also for her comments on an earlier draft version of this chapter.

[1] See C Chan and others, 'Artificial Intelligence Applications in Financial Services – Asset Management, Banking and Insurance' (*Oliver Wyman Research Report*, 2019), www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/dec/ai-app-in-fs.pdf; for an overview, T Boobier, *AI and the Future of Banking* (2020), and also T Guida, *Big Data and Machine Learning in Quantitative Investment* (2019) (hereafter Guida, *Big Data*) for more recent developments in quantitative investment.

[2] See D Zetzsche and others, 'Artificial Intelligence in Finance – Putting the Human in the Loop' (2020) University of Hong Kong Faculty of Law Research Paper No. 2020/006 (hereafter Zetzsche and others, 'Artificial Intelligence in Finance'), Guida, Big Data (n 1), or the recent regulatory proposals from the Monetary Authority of Singapore (2019).

governance framework for financial institutions.[3] The EU also explored this topic and published a report on big data risks for the financial sector, including AI, stressing appropriate control and monitoring mechanisms.[4] Scholars have developed this topic further by adopting so-called personal responsibility frameworks to regulate any new emerging AI-based applications in the financial industry.[5] In its recent draft regulation, the EU has presented a general risk-based regulatory approach of AI which regulates and even prohibits certain so-called high risk AI system; and some of them can supposedly also be found in the financial industry.[6]

This chapter will explore this entire topic of AI in the financial industry (which will also be referred to as robo-finance) further. One focus of the article will be on whether AI in the financial industry gives rise to new kinds of risks or merely increases existing risks already present in the industry. Further, the article will review one prominent general regulatory approach many scholars and regulators have put forward to limit or mitigate these alleged new risks, namely the so-called (personal) responsibility frameworks. In the final section of this chapter, a different proposal will be presented on how and to what extent best to regulate robo-finance, which will take up key elements and concepts from the recent Draft EU AIA.[7] To lay the groundwork for the discussion of these topics, the nature of AI, in particular as a general-purpose technology, will be explored first. In addition, an overview of the current state of AI applications in financial services will be given, and the different regulatory layers or focus areas for regulations that are present in the financial industry today will be presented. Based on these introductory discussions, the main topics of the chapter can then be spelled out.

II. AI AS A NEW GENERAL PURPOSE TECHNOLOGY

Electricity is a technology or technology domain which came into life more than 150 years ago, and it still drives a lot of change today. It comprises different concepts like electrical current, electrical charge, electric field, electromagnetics etc. which have led to many different application areas in their own right; from the light ball to electrical telegraphs or to electric engines, to mention only a few. It is fair to say that electricity as a technology field or domain has revolutionized the world in many ways, and it still does. And it has changed and transformed whole industries as it transforms the automotive industry with the transition from combustion engines to electric cars.

Given its wide range of underlying concepts with multiple specific application areas of their own right, several authors have referred to electricity as a general-purpose technology (GPT).[8]

[3] See the so-called IAC (Individual Accountability) guidelines by the Monetary Authority of Singapore, MAS, 'Guidelines on Individual Accountability and Conduct' (MAS, 10 September 2020) www.mas.gov.sg/-/media/MAS/MPI/Guidelines/Guidelines-on-Individual-Accountability-and-Conduct.pdf.

[4] See the joint report of the European Supervisory Authorities EBA, ESMA, EIOPA on the use of big data, including AI, by financial institutions, December 2016, JC/2016/86.

[5] See for instance Zetzsche and others, 'Artificial Intelligence in Finance' (n 2).

[6] The EU published the General Regulation on a European Approach for Artificial Intelligence in 2021, which regulates the financial industry in some areas of AI applications as well, European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM (2021) 206 final (hereafter Draft EU AIA). Since this draft was published after the writing of this article, its impact will and can be discussed only to a smaller extent in this paper.

[7] See Zetzsche and others, 'Artificial Intelligence in Finance' (n 2).

[8] See E Brynjolfsson and A McAfee, 'The Business of Artificial Intelligence' (*Harvard Business Review*, 18 July 2017) https://hbr.org/2017/07/the-business-of-artificial-intelligence 3 (hereafter Brynjolfsson and McAfee, 'The Business of Artificial Intelligence'), see also the interview with *Andre Ng* in M Ford, *Architects of Intelligence – The Truth about AI from the People Building It* (2018) 190 (hereafter Ford, Architects of Intelligence)

What is characteristic of GPTs is that there exists a wide range of different use cases in different industries, thus GPTs are not use-case-specific or industry-specific technologies but have applications across industries and across many types of use cases. Other examples of GPTs which scholars have identified are the wheel, printing, the steam engine, and the combustion engine, to mention a few.[9] As such, GPTs are seen as technologies that can have a wide-ranging impact on an entire economy and, therefore, have the potential to drastically alter societies through their impact on economic and social structures.[10]

Several authors have claimed or argued in recent years that AI can or should also be considered a GPT, 'the most important one of our era' in fact.[11] Or as *Andrew Ng* says: 'AI will transform multiple industries'.[12] AI's impact on societies as a whole is seen as significant, for instance, changing the way we work, the way we interact with each other and with artificial devices, how we drive, how wars might be conducted, etc. Further, like in the case of electricity, there are many different concepts underlying AI today, from classical logic or rule-based AI to machine learning and deep learning based AI, as employed so successfully today in many areas. Some hybrid applications combine both concepts.[13] These concepts have allowed for many new types of AI applications, similar to the case of electricity, where different concepts have been merged together as well.

In fact, because the use cases for AI technologies are so enormous today, companies like Facebook have created their internal AI labs or what they have called their 'AI workshop' where many different applications of AI technologies, in particular machine learning applications, get explored and developed.[14] The underlying assumption of such companies is that AI can be applied to so many different areas and tasks that they need to find good ways to leverage their technological expertise in all such different areas.

Clearly, AI is still in its early stages of technological development, with fewer implementations in widespread operation than in the case of electricity. But there have been language and speech processing applications, visual recognition applications like face recognition in smartphones, photo optimization algorithms in digital cameras, many kinds of big data analytics applications, etc. AI technologies have also changed the interface between humans and machines, some turn machines into helpful assistants, others allow for intelligent ways of automating processes and so on. The applications of AI are already widespread today, and we seem to be just at the beginning of a long journey of bringing more applications to life.[15]

In the following, we will look at the financial industry as one major application area for AI as a general-purpose technology. The financial industry is interesting in so far as it is heavily regulated on the one hand, but also highly digitalized and technologically advanced on the other hand, with many kinds of AI use cases operational already today.

[9] See R Lipsey and IC Kenneth, *Economic Transformations: General Purpose Technologies and Long Term Economic Growth* (2005) (hereafter Lipsey and Kenneth, Economic Transformations) for a broader discussion of different GPTs and their role for economic development and the transformation of societies as a whole.

[10] Besides Lipsey and Kenneth, Economic Transformations (n 9) see also TF Bresnahan and M Trajtenberg, 'General Purpose Technologies "Engines of Growth"?' (1995) 65(1) *Journal of Econometrics* 83 for another interesting article on the wider topic of the role and impact of GPTs.

[11] See Brynjolfsson and McAfee, The Business of Artificial Intelligence (n 8) 4.

[12] See the interview with *Andrew Ng* in Ford, *Architects of Intelligence* (n 8) 190 *et seq*.

[13] See ibid.

[14] See J Candela and S Berinato, *Artificial Intelligence: Insights You Need from Harvard Business Review* (2019)

[15] It is worth noting that today it is not entirely clear which direction AI as a technology will go over the next years. Despite the enormous success of machine learning as an AI concept or paradigm, several authors have pointed to its limitations – see for example the interview with *Barbara Grosz* in Ford, *Architects of Intelligence* (n 8) 333–356.

III. ROBO-FINANCE: FROM AUTOMATION TO THE WIDE-SPREAD USE OF AI APPLICATIONS

The financial industry has been one of the most data-intensive and digitized industries for decades. In 1973, SWIFT was founded and launched, the so-called Society for Worldwide Interbank Financial Telecommunication, bringing together 239 banks from 15 countries worldwide with the aim of handling the communication of cross-border payments. The main components of the original system included a computer-based messaging platform and a standard message system.[16] This system disrupted the manual processes of the past, and today more than 11,000 financial institutions from more than 200 countries are connected through SWIFT's financial global technology infrastructure. Nasdaq, to give another example, the world's first electronic stock exchange, began its operations even earlier, in 1971, leading the way to fully digitized exchanges for the trading of any kinds of financial securities, which are the standard and norm today. And real-time financial market data and news, probably the first big data sets used in history, were made available in the early 1980s by companies such as Thomson Reuters and Bloomberg through their market data feeds and terminal services.[17]

In the years to follow, the financial industry has been at the forefront of leveraging information (management) systems to manage and process the vast amounts of data and information available.[18] In fact, today, many financial institutions resemble technology companies more than traditional banking houses, and it is no surprise that companies like Paypal or, more recently, many new fintech players were able to further transform this traditional industry by leveraging new technologies like the Internet or mobile services, platforms, and infrastructures.[19]

This development of digitizing financial information and financial transactions has made the automation of data handling and processing, not just a possibility but rather a necessity to maintain and defend one's competitiveness and to deal with and manage the various kinds of risks inherent in the financial industry. The execution of payments within the international banking system or the execution of buying or selling orders on the exchanges can be fully automized today based on simple parameters (such as dates and amounts, or stop-loss orders to manage risk, etc.). Clearly, these ways of automizing financial transactions and processes are in no way intelligent, nevertheless they have helped the investment banks and other actors in the financial industry tremendously to increase process speed, accuracy, and also improve

[16] See www.swift.com/about-us/history for more details on the introduction of the SWIFT system.

[17] Commonly big data sets are defined by the so-called 4 Vs: volume (the amount of data), velocity (the speed in which new data get created or are generated), variety (different kinds of data types from different data sources, in particular, often a mix of structured and unstructured data), and veracity (discrepancies, errors, and gaps in data sets). Typical market data feeds in the financial industry fulfill at least three of these criteria, namely volume, velocity and veracity, as the data feeds deliver fairly structured data sets. This might change in the future when data feeds might also include other kinds of data such as press releases or social media posts as it is the case already with so-called sentiment feeds including sentiment data. See B Marr, *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance* (2015) for a more general introduction in the area of big data, and Guida, *Big Data* (n 1) for more insights into big data in areas of financial information.

[18] In broader terms an information (management) system is simply defined as a set of interrelated components consisting of an application system, and an interface for human interaction to define the tasks for the system and retrieve information. The application system consists of hardware, software, data, and a network connection. For more details see K Laudon and J Laudon, *Management Information Systems: Managing the Digial Firm* (15th ed. 2018).

[19] Examples in the payment sector are WeChat Pay, Alipay or Apple Pay, new competitors to the established credit card payment services. In fact, today many big tech companies are moving into financial services with their own finance applications, often in areas like payments, as Apple Pay or Google Pay.

risk management.[20] Therefore, it is no surprise that many actors in the industry have constantly searched and tried to develop more sophisticated processes, which has opened the doors for AI applications in the financial industry.

Today, there is a wide range of AI applications present in the financial industry of which the following are just key application areas with multiple kinds of use cases:[21]

(1) *Customer Related Processes:*
   a. new ways of segmenting customers based on the use of so-called cluster algorithms or analyses,[22]
   b. personalized banking services and offers based, for instance, on profiling algorithms,[23]
   c. robo-advisory services replacing human financial advisory with machines,[24]
   d. intelligent chatbots advising or providing information to clients in different areas of their financial decision making.[25]

(2) *Operations and Risk Management:*
   a. underwriting automation in credit decisions and algorithmic credit scoring,[26]
   b. automized stress testing.

(3) *Trading and Investment Management:*
   a. algorithmic trading – from simple rule-based AI to more sophisticated machine learning based algorithms,[27]
   b. automatic portfolio rebalancing in asset management adjusting the portfolio to the predefined asset allocation scheme based on simple rule-based algorithms,

[20] They operate more like a thermostat for a heating system, setting thresholds for certain actions to take place, like selling a stock position based on a predefined stop-loss order. The system will automatically initiate the transaction, but it is solely based on predefined parameters.

[21] Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) present a similar classification of the AI application present today in the financial industry. See also T Boobier, *AI and the Future of Banking* (2020). For discussion of several of the application areas discussed here, as well as a recent leadership paper by the consultancy firms Oliver Wyman, Marsch, BCLP and Hermes, C Chan, and others, 'Artificial Intelligence Applications in Financial Services – Asset Management, Banking and Insurance' (Oliver Wyman Research Report, 2019), www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/dec/ai-app-in-fs.pdf.

[22] See M Hassan and M Tabasum, 'Customer Profiling and Segmentation in Retail Banks Using Data Mining Techniques' (2018) 9(4) *International Journal of Advanced Research in Computer Science*.

[23] See R Ragotan, 'AI Has Changed the Way Banks Interact with Their Customers' (*Fintech News*, 5 February 2020) www.fintechnews.org/ai-has-changed-the-way-banks-interact-with-their-customers. For a discussion of some of the applications and service providers.

[24] So-called robo-advisors or advisory solutions like Betterment, Wealthfront, and Vanguard Digital, to give a few examples from the more advanced US robo advisory market, have in recent years been launched in competition with traditional human banking or financial advisors. These solutions automize and digitalize the advisory process in wealth management and private banking, thereby lowering the asset under management threshold for private investors for accessing high quality advisory solutions. Although some of the new players have also automized the asset management process itself, the primary focus of these solutions is enhancing the advisory process by replacing the human banking advisor with a machine or AI-based interface. In this regard they are classified here under customer related solutions and not under AI-based trading and portfolio management solutions as done by Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) which is rather misleading.

[25] As pointed out in a study by the consulting firm McKinsey & Company (2018), data analytics applications often using AI techniques are most widespread in sales and marketing areas of businesses, that is, areas which try to generate and develop new customer relationships and transaction.

[26] See an interesting article by N Aggarval, 'The Norms of Algorithmic Credit Scoring' (2021) 8(2) *The Cambridge Law Journal* 42 on the norms of algorithmic credit scoring.

[27] See M Lewis, *Flash Boys: A Wall Street Revolt* (2014) (hereafter Lewis, *Flash Boys*) or S Patterson, *Dark Pools: The Rise of the Machine Traders and the Rigging of the U.S. Stock Market* (2012) (hereafter Patterson, *Dark Pools*) for good non-expert introductions into this area, for a more systematic and scientific account see R Kissell, *Algorithmic Trading Methods: Applications Using Advanced Statistics, Optimization, and Machine Learning Techniques* (2021).

c. big data and machine learning–based (assisted or fully automized) asset management.[28]

(4) *Payment Processes:*
fraud detection algorithms in credit card payments using big data analytics and learning algorithms.[29]

(5) *Data Security and Cybersecurity:*[30]
a. data security – algorithms protecting the data from inside a financial institution,
b. cybersecurity – algorithms protecting the data from outside attacks.[31]

(6) *General Regulatory Services and Compliance Requirements:*[32]
a. Anti Money Laundering (AML) automation and protection algorithms helping to identify politically exposed people (so-called Peps) or criminals involved in certain financial transactions,
b. detection of compliance breaches in case of insider trading etc.

As shown here, AI is already employed today in many areas of the financial industry, and new applications are emerging every day. The question is whether additional or increased risks stem from these applications, which might require additional regulations, as argued by some authors.[33] This line of argument will be reviewed in more detail in the following sections. But first, it is important to understand from a high-level perspective the main areas and layers of regulations in the financial industry today.

IV. A SHORT OVERVIEW OF REGULATION IN THE FINANCIAL SERVICES INDUSTRY

The financial services industry is probably one of the most regulated industries. Regulation of trading practices for instance dates back to the seventeenth century when in 1610 in Holland, some first forms of *short selling* became prohibited.[34] At the same time, the first central banks were created, such as the Swedish Riksbank in 1668, to regulate payment transactions on a national level and establish national currencies by issuing banking notes. Some of the early

[28] See Guida, *Big Data* (n 1) on a recent collection of articles on this new emerging and developing field. So far, AI applications and tools have mainly been used in assisting fund managers in the asset allocation process, but it is possible that there will be fully AI-based fund management in the future. Some authors like E Syrotyuk, 'State of Machine Learning Applications in Investment Management' in T Guida (ed), *Big Data and Machine Learning in Quantitative Investment* (2019) seem to be more sceptical in regard to fully automized asset management because of the more erratic nature of financial markets.

[29] Companies like Teradata (teradata.com) and Datavisor (datavisor.com) provide AI-based financial fraud detection solutions. Datavisor, for instance, claims that their solution can detect 30% more frauds with 90% accuracy. Their solutions are mainly based on machine learning algorithms according to own research.

[30] See A Bouveret, 'Cyber Risk for the Financial Sector: A Framework for Quantitative Assessment' (2018) International Monetary Fund Working Paper 18/143 for a thorough overview and analysis of cyber security risk in the financial industry by sectors and countries/regions.

[31] See J Li, 'Cyber Security Meets Artificial Intelligence: a Survey' (2018) 19 Frontiers of Information Technology & Electronic Engineering for a more detailed analysis of the potential of using AI systems in preventing or reducing cyberattacks. The article also highlights the fact that AI systems might be used in facilitating cyber security attacks, as will be discussed also later in the article.

[32] A new sector has emerged in recent years often referred to as RegTech – see Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) – using technology to help financial institutions to comply with the various regulatory requirements. Quite a few regtech solutions have increasingly made use of AI technologies; for a good overview see 'AI in RegTech: a quiet upheaval' (*Chartis*, 2018) www.ibm.com/downloads/cas/NAJXEKE6.

[33] See Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) as a recent example.

[34] See AM Fleckner, 'Regulating Trading Practices' in N Moloney, E Ferran, and J Payne (eds), *The Oxford Handbook of Financial Regulation* (2015) 597 (hereafter Fleckner, 'Regulating Trading Practices').

regulation was 'private self-regulation', in other words, bottom up norm creation,[35] as in the case of regulatory practices around many of the emerging exchanges, but some regulation was already at these early times government- or state-driven (top down) as in the case of the establishment of central banks and their key role in establishing standardized payments practices based on backed up currencies.[36]

Today the financial industry is heavily regulated by national or supranational bodies, for instance, by the ESMA[37] in the EU or by the SEC[38] in the US in regard to activities on the different financial markets. Some of the regulations are financial-industry-specific, others are general regulations that severely impact the financial industry. Overall, the different types or layers of regulations in the financial industry can be classified by their underlying aims, namely: (i) regulations meant to safeguard overall financial stability, (ii) regulations for the protection of consumers of financial services, and (iii) regulations that are meant to make sure financial services can operate in a challenging and diverse international environment with sometimes conflicting rules and principles.[39]

The following overview tries to capture the main regulation areas or layers and their specific purpose or aim as they are present in the financial industry today. Some of the layers directly link up to the categories just mentioned, some are cutting across the different categories, and some are also mirroring the classification of the previous Section of AI-impacted application domains in the financial industry:

(1) Equity and liquidity requirements for banks and financial institutions to adhere to minimum capital ratios and liquid asset holdings to prevent financial stress, improve risk management, and promote transparency. Examples are the Basel I, Basel II, Basel III regulations which are global voluntary regulatory frameworks adhered to by most financial institutions today;[40]
(2) Infrastructure regulations, many still in the proposal stage, to improve financial services firms' operational resilience (in case of major disasters, for instance), and their responses to cyberattacks;[41]
(3) Pre- and post-trading regulations to strengthen investor protection and improve the functioning of financial markets, making them more efficient, resilient, and transparent like banning certain trading practices or making kickbacks by product issuers transparent. The MiFID I and II regulations in the EU are examples of such kinds of regulations;[42]

[35] For the different meanings of the notion 'regulation', cf. T Schmidt and S Voeneky, Chapter 8, in this volume.
[36] For a thorough analysis of regulation of trading practices in the financial industry discussing both sides of regulation see the article by Fleckner, 'Regulating Trading Practices' (n 34).
[37] European Securities and Market Authority (ESMA), the EU's securities market regulator located in Paris, created in 2011 and replacing the Committee of European Securities Regulators (CESR).
[38] Securities and Exchange Commission (SEC), the independent agency of the US federal government, created in the early 1930s following the stock market crash in 1929.
[39] See the KPMG report, 'EU Financial Services Regulation – A New Agenda Demands a New Approach' www.kpmg.com/regulatorychallenges, for giving a good overview on the various regulatory perspectives of regulation of financial services in the EU.
[40] See for a concise and high-level summary of the Basel I–III regulations the article 'History of the Basel Committee' (*BIS*) bis.org/bcbs/history.htm.
[41] See the European Commission, 'Proposal for the Regulation of the European Parliament and of the Council on Digital Operational Resilience for the Financial Sectors' COM (2020) 595.
[42] See M Comana, D Previtali, and L Bellardini, *The MiFID II Framework: How the Standards Are Reshaping the Investment Industry* (2019) for a detailed analysis of the MiFID II regulations including a comparison with the MiFID I rules.

(4) Payment services regulations, like the PSD II directive (2015) in the EU, with the aim of creating more integrated payments markets, making payments safer and more secure and also protecting consumers, for instance, from the financial damage resulting from fraudulent credit card payments;
(5) Various kinds of compliance regulations, for instance anti-money-laundering or terrorist financing regulations etc., to ensure that financial institutions obey the treaties and laws and do not enter any illegal transactions or practices, also regarding cross-border transactions;[43]
(6) *General data privacy protections* like the GDPR[44] in the EU, which is highly relevant as financial transactions involve much sensitive personal data.

As we can see, there are no specific AI regulations of financial services – although many of the regulations will also impact AI-based financial services. In fact, there are even very few regulations regarding the underlying technologies in financial services, but most of them focus on the use cases or financial activities, processes, and on the outcomes themselves. Yet, recently some scholars and some regulators have argued that there might be new risks stemming from AI applications and technologies in financial services which require additional regulation. In the following, we will look at some of the alleged risks as pointed out by scholars in the field and explore to what extent they might be covered by the above regulations already or whether there is a need for new regulatory frameworks.

V. NEW RISK CATEGORIES IN ROBO-FINANCE STEMMING FROM AI?

Dirk Zetzsche and others, in their recent paper, have identified the following four risk categories or risk areas allegedly related to AI applications in the financial industry:

(1) Data risks
(2) Cybersecurity risks
(3) Financial stability risks
(4) Ethical risks.[45]

Although I agree with the authors that all these kinds of risks are related to AI applications in financial services, it appears that these risks already existed before the emergence of robo-finance, given the advanced stage of the industry in terms of digitization and data dependency and usage. In fact, some of these risks might even be reduced or vanish when AI comes into place. Let us look at the different risk areas one by one.

Firstly, starting with the data risks of AI applications, *Zetzsche* and others bring up the following more specific arguments: (i) Because the data quality might be poor, there can be deficiencies stemming from AI applications. As a matter of fact, data quality has often been poor in many parts of the financial industry, for instance, outages at the data centers of the exchanges or of the market data providers leading to the misstating of prices of securities, which can have

[43] The laws and regulations around data privacy protections can also be seen as falling into this category but it has been listed here separately given its recent prominence.
[44] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.
[45] This classification mirrors or reflects to some extent on the classification by the French prudential regulatory authority within the bank of France. It has recently put forward the following four risk categories as allegedly stemming from AI applications in the financial industry: (1) data processing risk, (2) cybersecurity risk, (3) challenges to financial stability, (4) player's dependency and change in power relationships in the financial market. See 'Artificial Intelligence: Challenges for the Financial Sector' (*ACPR*, December 2018), acpr.banque-france.fr/sites/default/files/medias/documents/2018_12_20_intelligence_artificielle_en.pdf.

negative effects on investors' decisions and the markets overall. AI could actually be used to deal with the data issues in terms of detecting and even resolving them.[46] (ii) Besides, they argue that data used for AI analyses might suffer from biases, for instance, relating to what they call 'oversight' in a financial organization. Again, biases have already been influencing decision making in the financial industry even before the emergence of AI applications, maybe not in the form of what they call data-biases, but biases residing more generally in human decision, for instance in the making of credit decisions or consumer lending.[47] AI application might free us from certain biases by providing a more neutral stance if programmed accordingly or at least be sensitive to such kinds of biases. (iii) And it is claimed that AI interdependency can lead to what they call 'herding', for instance all systems selling securities triggered by certain market events, which can lead to what has been referred to as 'flash crashes'.[48] Again 'herding' behavior has existed in the financial markets for a long time, and whether the emergence of AI in electronic trading systems has been the cause of what has been called 'flash crashes' seems rather questionable. Simple rule-based algorithms, which today by industry experts would rather not be classified as AI systems can give rise to such behavior in contrast to more sophisticated systems trained on historical data relating to such events.

Secondly, let us look at cybersecurity risks. Obviously, they have also existed before the arrival of AI, with most attacks initiated and conducted by human individuals directly or by simple processes, methods, or algorithms. Examples are emails carrying malware that, after it has installed itself on someone's computer, can silently send all sorts of confidential data from the computer or computer network to the attacker; a similar case of phishing attacks through links to websites – for instance, of online banks that mimic the log-in pages one is familiar with; or finally, simply the reuse of a user's credentials which the attackers have somehow got hold of – for instance, by one of the already mentioned measures or by simply spying on people in combination with our carelessness in setting passwords. That 'algorithms can be manipulated in an effort to transfer wealth' has nothing to do with the presence of AI systems because this could be done already before such systems were in place and it currently happens every day in many different ways within traditional information system environments.[49] It rather seems plausible that AI might provide some help in identifying and preventing cybersecurity attacks.[50]

[46] For instance, the construction of error correction codes can be used in handling issues in data transmission through noisy channels as for instance happens sometimes in the case of market data feeds. More recently AI techniques have been used in optimizing the design of error correction codes, see for instance L Huang and others, 'AI Coding: Learning to Construct Error Correction Code' (2019) 20(10) *IEEE Transactions on Communications* (hereafter Huang and others, 'AI Coding').

[47] In their interesting paper W Dobbie and others, 'Measuring Bias in Consumer Lending' (2021) *The Review of Economics Studies* https://doi.org/10.1093/restud/rdaa078, tried to measure the amount of bias in consumer lending decision. What they found is that in traditional non-AI-based lending decisions there is a significant bias against immigrant and older loan applicants.

[48] There has been a lot of debate around the so-called flash crash which happened on May 6, 2010, when the Dow Jones Index lost about a tenth of its value in just 36 minutes – see for instance A Kirilenko and others, 'The Flash Crash: The Impact on High Frequency Trading on an Electronic Market' (2017) 72 *The Journal of Finance* 967. In his recent article D Busch, 'MiFID II: Regulating High Frequency Trading, other Forms of Algorithmic Trading and Direct Electronic Market Access (2017) 2 *Law and Financial Markets Review* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3068104 (hereafter Busch, 'MiFID II') looks at how, by the MiFID II regulation, such flash crashes are meant to be banned by ruling out a technique of market manipulation referred to as 'spoofing'. This technique was allegedly used by a British stock market trader in 2010 when he tricked the market into believing that the prices were about to fall by placing huge amounts of sell orders which were later cancelled by him by his specially developed algorithms.

[49] See Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) 21.

[50] See the discussion in Section II (5).

Many services are offered today in this regard, and this seems to be one of the areas where the financial industry could benefit from employing AI-based solutions and thereby reduce potentially harmful cybersecurity risks.

Thirdly, when *Zetzsche* and others talk about financial stability risks, it is fairly unclear what they have in mind since they mention almost all areas of AI applications in financial services – as laid out above – from consumer facing and supporting applications, to trading and portfolio management systems, to general regulatory and compliance systems. Overall, their main concern here seems to be the emergence of 'additional third-party dependencies' to AI technology providers up to the 'level of oligopoly or monopoly'. Since many of these third-party technology providers are unregulated today, as they point out, and it might even be hard to regulate them as 'AI-related expertise beyond those developing the AI is limited', there appears to be a major risk. As they say, 'these third-party dependencies [...] could have systemic effects'.[51]

What can be said against this last part of their arguments is that many of the actors at the forefront of using AI in financial services today develop their applications inhouse, like the dominant hedge funds or algo trading shops set up by IT specialists.[52] AI-based technology has become such a core asset these days and a competitive factor of financial services that many financial institutions resemble IT companies more and more today to keep all that knowledge inhouse or with high IT expertise inside the organizations to manage their IT service providers or outsourcing partners – far from being entirely dependent or in the hands of monopolistic or oligopolistic structured IT providers.[53] Hence, their worry about systemic effects stemming from such dependencies seems to be overstated, at least in certain critical banking areas. Moreover, in many instances there are quite a few technology providers that offer similar services to the financial industries, for instance market data providers which increasingly have started to use AI technologies to organize and manage the quality of their market data feeds.[54] Financial institutions, at least in critical areas like trading, often make use of different providers at the same time, which also helps them to reduce their third-party-dependencies. Furthermore, with higher education AI or machine learning programs popping up at many educational institutions around the world, new graduates are also increasingly being educated and trained in these key areas. Thus, knowledge is building up quickly and will also be more widely available, reducing the fear of there being a kind of 'mystery science' only a few people have access to and can take advantage of.

Finally, let us focus on what *Zetzsche* and others refer to as new 'ethical risks' stemming from AI applications in financial services. The starting point of their argument is that algorithms do not feel anything, nor do they have values which the authors equate with a lack of ethical foundation in AI-decision making. For instance, they point out that such 'unethical' AI systems

[51] See Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) 21.

[52] See Lewis, *Flash Boys* (n 27) and Patterson, *Dark Pools* (n 27) for a vivid description of the individuals, often IT experts or nerds, in setting up high frequency trading firms or the respective trading units at major banks. Also, major hedge funds with a quantitative focus like Renaissance Technologies, which had 133 billion USD under management as of November 2020, have a strong focus on developing their own mathematical models and algorithms.

[53] See the article by T York, 'Banks Becoming Technology Companies, Technology Companies Becoming Banks' (*San Diego Business Journal*, 30 September 2019) www.sdbj.com/news/2019/sep/30/banks-becoming-technology-companies-technology-com/; see also the recent BCG publication on this topic, J Erlebach and others, 'The Sun Is Setting on Traditional Banking' (*BCG*, 24 November 2020) www.bcg.com/publications/2020/bionic-banking-may-be-the-future-of-banking.

[54] For instance, market data providers like Bloomberg and Thomson Reuters have started to use AI methods and techniques, helping to digest larger data sets including unstructured data like texts from different sources, thereby delivering new kinds of analytics such as so-called sentiment analysis or feeds, trying to identify the sentiments in certain markets or regarding certain securities.

might nudge people to purchase unsuitable financial products, which might further be facilitated by the fact humans would easily develop a higher level of trust in the AI-based systems because with them, human–machine communication can nowadays be quite sophisticated. What this ultimately can and will lead to is reputational risk for the financial institution employing such systems, for example, when people are driven to make the wrong financial decisions and this becomes public or will be reported in the media or brought up to the courts.

There are quite a few problems with this line of reasoning as there are many financial institutions that do not have much direct interaction with human consumers, like mutual funds, hedge funds, credit card companies, etc. Besides, it is also conceivable that AI systems can have an ethical foundation, for instance thinking of utilitarian approaches which are less focused on being able to feel anything or have values. Such aligned AI systems might still be able to calculate the best outcome for society as a whole. But the main counter argument seems to be that the financial industry has not been a role model for ethical behavior to start with. Quite to the contrary, over many decades, financial institutions have been prone to all kinds of ethical misconduct. Just to give a few examples: (i) consumers have been pushed by financial advisors, humans with feelings and values, employed by financial institutions, to buy financial products which were often not suitable or beneficial to them, yet by selling them, the advisors were able to boost their commission payments, and the financial institutions could thereby boost their profits;[55] (ii) insider trading has happened frequently,[56] (iii) market manipulation has occurred, for instance in the case of the Libor scandal, and many other examples in different areas of the financial industry.[57] Thus, it is far from clear why AI-based systems and processes would make the industry less ethical than it has been in the past. In fact, the case could be made that AI-based systems and processes might allow society to create and control financial institutions and make them less driven by greed but more by higher motives to bring benefits to consumers and install fairness within the systems.

But this line of reasoning might sound overly naïve, given how many actors in the financial industry have successfully used technology over the last decades to their advantage, and to the disadvantage of other actors. One example has been the area of high frequency trading and the so-called dark pools where 'fast moving robot trading machines were front-running long term investors on exchanges'.[58] Darks pools are markets established by the financial actors themselves

[55] In Germany for instance the advisory services offered mostly by banks have been reviewed frequently by consumer protection agencies and independent bodies, and over many years the findings have been very disappointing with many banks not even fulfilling basic standards and requirements – see the magazine *Finanztest* 2/2016. In particular, elderly people have been frequently 'ripped off' and have been referred to internally as 'AD's (alt (old) and dumm (stupid)), to whom the advisors could sell products not suitable to the financial situation of the elderly or asking them to re-allocate their portfolio frequently mainly with the aim of generating extra commission fees on the triggered transaction, thereby exploiting their trust – see C Bauer, 'Banken zocken Senioren als "AD-Kunden" ab' *Westfaelische Rundschau* (9 July 2009) www.wr.de/wr-info/banken-zocken-senioren-als-ad-kunden-ab-id79712.html. In States like the US, where there has been a long tradition of investing in the financial markets also by private investors through their 401K pension plans with tax benefits, financial advisory services have been on higher professional levels. For a more thorough cross-country comparison see J Burke and A Hang (2015), 'Financial Advice Markets – A Cross-Country Comparison' (study by the Rand Corporation prepared for the US department of labor) www.rand.org/pubs/research_reports/RR1269.html.

[56] There has been a long history of insider trading; see the article by the New York Times, 'Dealbook – Timeline: A History of Insider Trading' *The New York Times* (6 December 2016), mainly focusing on cases in the US www.nytimes.com/interactive/2016/12/06/business/dealbook/insider-trading-timeline.html.

[57] Over many years traders had manipulated the banks' central lending rate, i.e. the LIBOR rate, to their benefit before it was discovered, see L Vaughan and G Finch, 'Libor Scandal: The Bankers Who Fixed the World's Most Important Number' *The Guardian* (18 January 2017) www.theguardian.com/business/2017/jan/18/libor-scandal-the-bankers-who-fixed-the-worlds-most-important-number.

[58] See Patterson, *Dark Pools* (n 27) 4, and also M Lewis, *Flash Boys* (n 27) for more details on this fascinating topic.

for trading securities outside of the exchanges, usually virtually unregulated. The benefits for the market actors were faster processing of orders with less or even no fees from the exchanges. But the real benefits for the involved high frequency trading firms were obviously financial: with their algorithms and high frequency trading infrastructures, they were able to read the directions markets were going, and being able to buy securities before the real investors could do it and then selling the securities back to them at a higher price only milliseconds after the initial orders by the investors had been made. This practice allowed them to make huge profits stealing away money from the long-term investors like pension funds, etc.

This is a very sophisticated version of an old, mostly considered illegal practice of so-called front-running – in other words, someone trading a stock or any other financial asset based on insider knowledge of a future transaction that is about to affect its price. One important point is that this practice has been around before the emergence of AI and the high frequency data processing infrastructures. But it needs to be acknowledged that the new technologies have allowed for a more sophisticated and harder-to-control form of front-running. Yet the problem here is not that the employed AI algorithms are unethical. The problem is that the actors have used the technologies in an unethical way which obviously needs to be prevented for the benefits of the wider investor community and for society as a whole. Again, this is not a new risk, but it shows that AI and technology can be an accelerator of existing risks inherent in the financial industry.

In sum, then, the arguments by *Zetzsche* and others that there are many new risks stemming from AI-based applications in financial services are not fully convincing. To the contrary, it is feasible that the employment of AI applications in the financial industry might provide a route of managing existing and inherent risks in a better way, or even being able to reduce or eliminate some of these risks.[59] But clearly, there are also cases like front-running based on algorithmic high frequency trading, where it seems obvious that through the employment of AI, existing inherent risks in the financial industry have increased and can cause additional damage. Therefore, it is also important to look at ways how such damages resulting from the use of the new technologies can be avoided. In this regard, in the following section one prominent regulatory approach, the so-called responsibility frameworks, will be discussed.

VI. RESPONSIBILITY FRAMEWORKS AS A SOLUTION FOR MANAGING AI RISKS IN FINANCIAL SERVICES?

In regulating the financial industry, many regulators have moved to so-called responsibility frameworks in recent years, like the EU's EBA/ESMA guidelines or the FCA in the UK.[60] The proposed measures focus on personal managerial responsibility, for example, the personal responsibility of directors, senior management, and individual line managers. Initially, such frameworks were meant to be applied to mitigate the risks of financial services in general, but recently authors have argued that they can also be applied to the emerging AI-based processes in the financial industry.[61]

[59] For another view, cf. T Schmidt and S Voeneky, Chapter 8, in this volume.

[60] See the report by European Banking Authority for example: EBA, 'Final Report on Guidelines on internal governance under Directive 2013/36/EU' (2 July 2021) EBA/GL/2021/05, 5-7 www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Guidelines/2021/1016721/Final%20report%20on%20Guidelines%20on%20internal%20governance%20under%20CRD.pdf. For the UK, see the conduct rules as applied to the senior management functions as defined by the Bank of England report, Bank of England 'Senior Managers Regime: Approvals' www.bankofengland.co.uk/prudential-regulation/authorisations/senior-managers-regime-approvals.

[61] Cf. Zetzsche and others, 'Artificial Intelligence in Finance' (n 2).

The responsibility-driven regulations by the EU, published by EBA and ESMA, focus mainly on the management bodies of financial institutions, in particular on their role in conducting their overall operational duties, but with a particular focus on risk management conduct. They are meant to ensure that a sound risk culture has been implemented in their respective organizations consistent with the individual risk profile and the overall business model of the institution. The UK's senior management regulatory framework for financial institutions has evolved from the overall EU framework but it has strengthened the establishment of clear conduct rules for senior managers. These rules specify in more detail the steps necessary to ensure that the business of the financial institution is controlled effectively and is in compliance with existing regulatory frameworks. Also, requirements are made on the delegation of responsibilities and on the disclosure of relevant information for the regulators. Other States like the US or Singapore have issued similar guidelines.[62]

Although these responsibility frameworks have been very general in nature, meant to capture all kinds of aspects of risk management in financial institutions, Zetzsche has argued that they will give us the right framework to address and manage any new risks stemming from AI applications in financial services. They write: 'personal responsibility frameworks provide the basis for an appropriate system to address issues arising from AI in financial services'.[63] They have suggested the following three distinct instruments or measures for regulating activities related to the development and use of AI applications in financial services:

1. *AI Review Committees:* the installation of AI review committees is meant to address what they call the information asymmetry as to the function and limits of an AI system, namely the problem that third party vendors or inhouse AI developers understand the algorithms far better than the financial institutions that acquire and use them, and the supervisors of the institutions. These committees are meant to augment decision making and should not 'detract from the ultimate responsibility vested in management [. . .] regarding AI governance.[64]
2. *AI Due Diligence:* mandatory AI due diligence should be put in place, which should be done prior to any AI employment and should include what they call "a full stock of all the characteristics of the AI [. . .] in particular the mapping of the data set used by AI", including an analysis of data gaps and data quality.[65]
3. *AI Explainability:* the explainability requirement is proposed to be necessary as a minimum standard 'demanding that the function, limits and risks of AI can be explained to someone at a level of granularity that enables remanufacturing of the code'. And this someone 'should be a member of the executive board responsible for the AI'.[66]

Before we review this proposal, it is fair to mention that the authors themselves note a few limitations, of which I want to focus on the main one, namely the inability of their responsibility framework to control what they call 'autonomous AI'. What they mean by this are cases in which developers lose control over self-learning AI, not understanding anymore what the algorithms are doing.[67] What they propose is the concept of being able to always switch off the AI (as a kind of

[62] MAS, 'Guidelines on Individual Accountability and Conduct' (MAS, 10 September 2020) www.mas.gov.sg/-/media/MAS/MPI/Guidelines/Guidelines-on-Individual-Accountability-and-Conduct.pdf, para 3.3.
[63] Zetzsche and others, 'Artificial Intelligence in Finance' (n 2) 44.
[64] Ibid..
[65] Ibid.
[66] Ibid.
[67] Such a situation seems not so rare as also discussed in the recent documentary 'The Social Dilemma' (2020) on Netflix, in which many of the creators of the algorithms underlying the leading social media platforms like Facbook or Youtube discuss their inability to understand the content proposing aspects based on user profiling at a later stage of

human oversight) while the provided services would still be functioning. This seems, *prima facie*, to be a reasonable request, looking for instance at the example of self-learning AI application in payment fraud detection based on the analysis of large transaction data sets. The system might modify its outlier detection algorithm in a way which might force the financial institution to switch it off, maybe because fraudsters have fed the system with data to facilitate fraudulent transactions. In this setting, switching the AI system off would make sense, but the delivery of the basic payment services should not be impacted by this – for instance, in the case of a credit card company. Yet, there will be applications where such a switch-off mechanism might be more difficult to realize without causing any further damage, as in the case of trading financial securities where orders or transactions 'might get lost' by switching off applications.[68]

Overall, I agree with the authors that their approach is important and should be part of any software-based technology development in financial services. In fact, many elements have been in place in the industry for years already, for instance in terms of regular due diligence audits of the financial services' technology providers.[69] Hence, the financial industry is already prepared and experienced in conducting due diligence audits on a regular basis, and they do this frequently before the release of new technology and software systems or installations, irrespective of whether these systems would include AI technology or not.

Yet, on the other hand, the authors propose some specific requirements for their AI due diligence and also in regard to their explainability requirement for AI. As will be argued, these requirements are not entirely clear and potentially will also be hard to fulfill given the nature of many AI applications.

Firstly, they argue that any AI due diligence should comprise taking full stock of all the characteristics of the AI, 'in particular the mapping of the data set used by AI, including an analysis of data gaps and data quality. It is not clear what is meant by 'a full stock of all the characteristics of the AI'. Besides talking about the different functionalities of the AI system, they also seem to focus on the underlying data set used by it. The problem here is that data sets might be potentially infinite and/or not be fully determined at the outset. In the case of its employment, a self-learning AI might discover new data (sets), and as it is often the case with big data, there can be quality issues and gaps. What does this mean for the AI system: should it not be launched under such circumstances? Or is it just enough to be aware of such limitations?

Secondly, their explainability requirement seems even harder to deal with in the case of AI applications. Even in regard to existing non-AI applications, it is questionable whether this requirement can be met given the complexity of many software solutions in the financial sector with millions of lines of code and often old legacy systems.[70] In the case of AI-based application, the situation is even more complex because learning AI systems are less static but more dynamic in nature, which could mean that the system might even rewrite its code in the course of its operations. Making explainability a minimum standard in the sense defined above could be the

the operations of the algorithms. Essentially, the algorithms develop in their own way, which is hard to understand at later stages of their employment.

[68] The other two limitations they mention are overdeterrence – as long as the benefits are higher, I think this won't be such an issue – and the increased role of fintechs in developing AI applications which usually have less experienced managerial resources. Here, they propose that by suitable board structures this could be handled, a thought with which I agree.

[69] For instance, regarding the numerous cloud-based services in place today in areas like financial market data systems, trading terminals, or wealth management advisory solutions.

[70] See for instance the recent 2020 report by the consultancy firm Deloitte on this topic: 'Modernizing Legacy Banking Systems Practical Advice to Help Banks Succeed at Core and Application Modernization' (*Deloitte*, 2020) www2.deloitte.com/us/en/pages/financial-services/articles/modernizing-legacy-systems-in-banking.html.

end of many applications in financial services, not only AI-driven ones. But what might be argued for is that a reduced version of this principle can be applied, not to require the remanufacturing of the actual code but, at a minimum, the possibility for the functions, limits, and risks of the AI systems to be understood on a higher functional level.

Thus, to conclude this section, the so-called responsibility frameworks can provide a basis for limiting the risks of AI applications in financial services, given that new risks really emerge. In a way, they have been present in the financial industry before the rise of AI applications, such as regular due diligence audits of technical systems and key software applications. But they also have their limitations, in particular, when certain requirements are taken in a very strict sense, as it was the case above in regard to the discussed explainability requirement. Yet reducing such requirements to a lower level raises the question to what extent the risk stemming from any kinds of new potential high risk AI applications in financial services can be contained. A slightly different approach will be presented in the final section, which builds on the approach put forward recently by the EU.

VII. STANDARDIZATION, HIGH RISK AI APPLICATIONS, AND THE NEW EU AI REGULATION

As noted before in this chapter, AI, being a general-purpose technology, will impact not just one industry but will also have many different kinds of use cases in different industries. There are already a lot of AI applications in the financial industry today, as pointed out earlier, and more are being added constantly by different financial institutions, their IT service providers, and innovative fintech companies. Many of the solutions might be simple or fairly basic, like the use of face recognition as an identification method giving users access to their financial accounts or applications. The others will be more complex, like developing so-called robo advisors, with AI systems engaging with users in natural languages trying to understand their financial needs and giving them suitable financial advice.

What will be important going forward is that regarding some kinds of key AI applications – such as identification processes or human–machine interaction – there can and will be standards across industries which companies need to comply with, like there are standards and norms for the use of electricity irrespective of their specific domain of use.[71] Looking at the example of AI systems intended to interact with humans, providers of such systems might require that they be transparent to the user, that they are not communicating with humans but with a machine. Such notification obligations can then become part of the standard for such human–machine interaction enabling AI systems.[72]

Besides such general standard AI applications used across industries, there might also be ones very specific to certain industries like the financial industry which need to be dealt with outside of the model of standardization. In particular, when these specific applications give rise to higher or new risks, additional specific regulations might need to be put into place. For instance, there have been attempts to contain the risks of algorithmic trading applications in the financial industry, which can cause (and probably have already caused to some extent) significant financial damage in the form of leading markets to crash, thereby diminishing or blowing away

[71] See for instance all the different norms and standards defined by the VDE (the German association for electrical, electronic, and information technologies) over more than 100 years. In 1885, the first VDE regulation, the 'VDE 0100', was introduced, which regulated the safe construction of electrical systems. In 1904, the VDE published its first 'book for standards' comprising more than 17 provisions. Today, there exists a wide group of norms and standards ensuring the safety and well-functioning of all kinds of electrical systems.

[72] A similar obligation has been put forward by the EU in its recent Draft EU AIA, (n 6).

investors' money in the course of seconds.[73] Such flash crashes have been at the center of some debates over the last years, and regulators such as the EU have tried to contain this risk by putting additional obligations into place through the MiFID II framework as discussed earlier.[74]

In its recent Draft EU AIA the EU has also distinguished between 'high risk' and non-high-risk (standard) AI applications.[75] The new proposed regulation starts with the assumption that AI applications are ultimately and potentially just tools to increase human well-being. Thus, the technology development of AI should not be hindered by any unnecessary constraints, but the rules should be balanced and proportionate. The regulation is centered on a 'risk based regulatory approach, whereby legal intervention should be tailored to those concrete situations where there is a justified cause of concern'.[76] A key distinction is made between the so-called high risk AI systems for which special requirements and obligations then apply and other AI systems with much more limited requirements and obligations. The classification of AI systems as high-risk is thereby mainly based on their intended purpose and their harmful impact on health and safety and human rights. High risk systems are more or less identified in a two-step process, namely whether they can cause certain harms to protected goods or rights and by the severity of the harm caused and the probability of occurrence.

Given this approach, it seems obvious that there cannot be one final list of high-risk AI applications because the technology is still emerging and new applications are being launched every day. The EU acknowledges this as it lists in its Draft EU AIA only a limited number of high-risk AI applications (Annex III). Further, it allows the EU Commission to amend this list over time based on criteria spelled out in Article 7.

Interestingly, many of the high-risk applications listed by the Draft EU AIA are not specific to one industry but are general AI applications that can be present in many industries. Examples are applications that embody what is called 'manipulative AI practices' and a second group with 'indiscriminate surveillance' practices. But there are also many very specific high risk AI applications listed in the draft. When it comes to high-risk AI applications in financial services, the EU draft of the regulation lists, prima facie, only one class, namely AI systems that evaluate the creditworthiness of persons (Annex III No 5 lit. b).[77] This class of applications is included in the high-risk list because of (i) possible discrimination of persons of certain ethnic or racial origin based on the potential perpetuation of historical patterns by the AI algorithms, and (ii) the potential severity of such acts of discrimination, as in the way such discriminating credit decisions can significantly affect the course of life of people.[78]

The second kind of AI application that can be associated with the financial services industry, listed in the Draft EU AIA, is the one written about above, namely AI systems intended to

[73] In the literature, there has been a long discussion of the so-called flash-crashes and the extent to which they have been caused by certain algorithmic trading practices. See Busch, 'MiFID II' (n 48) on this topic for a more detailed discussion regarding the recent MiFID II regulation and its impact on algorithmic trading practices. See also Huang and others, 'AI Coding' (n 46) for more details on this topic.

[74] For more details on this topic see Section IV and Huang and others, 'AI Coding' (n 46) of this chapter.

[75] For details cf. T Burri, Chapter 7, T Schmidt and S Voeneky, Chapter 8, and C Wendehorst, Chapter 12, in this volume.

[76] See the Draft EU AIA, (n 6).

[77] For requirements to be met by high-risk AI systems, cf. Article 8 et seq., Article 16 et seq., and especially the conformity assessment, Article 43.

[78] I assume this refers to the fact that simple learning algorithms might be trained on past credit decisions of financial institutions which might have embodied certain forms of discrimination. As has been pointed out before in Section IV, also before the arrival of AI in financial services, many credit decisions have been prone to discrimination. One solution could be that in training algorithms on making such credit decisions the training data could be prepared in a way that would make them bias free.

interact with natural persons or generate content consumed by such person. Such systems do not necessarily classify as high-risk systems, for instance they might just help someone to enter information or explain a product, but they pose the specific risk of impersonation and deception, and therefore they are subject to specific transparency obligations according to the Draft EU AIA that means that the natural persons have to be informed that they are interacting with an AI system (Article 52).

Overall, the risk-based regulatory approach which underpins the Draft EU AIA makes much more sense than any kind of generalized approach of regulating AI applications as a whole as embodied in the responsibility frameworks discussed above. As a general-purpose technology, there will be so many kinds of applications that not one standard set of rules can be applied across the board. A rigorous case by case approach is required, which also allows for amendments and revisions, as embodied in the outline of the EU regulation.

VIII. CONCLUSION

What has been shown in this chapter is first, that it is questionable whether there are many new additional risks stemming from AI applications in financial services today. The risks that have emerged recently, like data risks, cybersecurity risks, financial stability risks, and ethical risks have been inherent in the financial industry as a highly digitized and also complex global industry for decades. The author has taken the more positive view that by using AI these risks will not necessarily increase, but on the contrary, AI might help to mitigate and reduce them. Second, the responsibility frameworks as developed over the last few years, which are meant to deal with and limit the risks of AI in the financial industry overall, do not provide a suitable framework beyond what has been put in place already to manage the risks with more standard IT and software systems and applications in the financial industry. Furthermore, overseeing all AI applications in financial services will quickly become as complex as overseeing all types of applications in the area of electricity, to mention another general-purpose technology. What has been argued in this chapter is that for some kinds of key applications – like identification processes or human–machine interaction – there should be standards defined across industries with which companies need to comply. But for other very specific, potentially new high-risk financial AI applications, in case they emerge, there might be the need for additional very specific regulation, as in the case of certain algorithmic high frequency trading applications. But this will be less a regulation of the technology but more of the practices and intended uses of the technology, which has also been the core thinking underlying the recent Draft EU AIA of AI applications. In fact, this new EU regulation, like the GDPR a few years ago in regard to data privacy protection, in many ways points to the right direction of how to deal with AI and potential risks arising from it.

PART VII

Responsible AI Healthcare and Neurotechnology Governance

22

Medical AI

Key Elements at the International Level

*Fruzsina Molnár-Gábor and Johanne Giesecke**

I. INTRODUCTION

It is impossible to imagine biomedicine today without Artificial Intelligence (AI). On the one hand, its application is grounded in its integration into scientific research. With AI methods moving into cancer biology, for example, it is now possible to better understand how drugs or gene variants might affect the spread of tumours in the body.[1] In genomics, AI has helped to decipher genetic instructions and, in doing so, to reveal rules of gene regulation.[2] A major driving force for the application of AI methods and particularly of deep learning in biomedical research has been the explosive growth of life-sciences data prominently based on gene-sequencing technologies, paired with the rapid generation of complex imaging data, producing tera- and petabytes of information. To better understand the contribution of genetic variation and alteration on human health, pooling large datasets and providing access to them are key for identifying connections between genetic variants and pathological phenotypes. This is not only true for rare diseases or molecularly characterized cancer entities, but also plays a central role in the study of the genetic influence of common diseases. The sheer growth and combination of data sets for analysis has created an emerging need to mine them faster than purely manual approaches are able to.[3]

On the other hand, based on this knowledge from biomedical research, the use of AI is already widespread at various levels in healthcare. These applications can help in the prevention of infectious diseases, for example by making it easier to identify whether a patient exhibiting potential early COVID-19 symptoms has the virus even before they have returned a positive test.[4]

* The authors acknowledge funding by the Volkswagen Foundation, grant No. 95827. The state of the science is reflected in this chapter until the end of March 2021. The sources have been updated until mid-September 2021.

[1] E Landhuis, 'Deep Learning Takes on Tumours' (2020) 580 *Nature* 550.

[2] Ž Avsec and others, 'Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax' (2021) 53 *Nat Genet* 354.

[3] E Landhuis, 'Deep Learning Takes on Tumours?' (2020) 580 *Nature* 550.

[4] S Porter, 'AI Database Used to Improve Treatment of UK COVID-19 Patients' (*Healthcare IT News*, 20 January 2021) www.healthcareitnews.com/news/emea/ai-database-used-improve-treatment-uk-covid-19-patients?utm_campaign=Clips&utm_medium=email&_hsmi=108004999&_hsenc=p2ANqtz-_Z0v3NgnQwS4wQHlc_eXjnWIuszmpfIvLSXXOM4z23_6DfT02WdoeI808wYaIICunIyDMs7g82wwC8V217XJn9K1SfCJByZihVAdmrIAS0yq7u7lt2Wd-d3JmQ5wwVPj09XOkf&utm_content=108004999&utm_source=hs_email; concerning the usefulness of AI applications for pandemic response, see: M van der Schaar and others, 'How Artificial Intelligence and Machine Learning Can Help Healthcare Systems Respond to COVID-19' (2021) 110 *Mach Learn* 1.

It can also help to understand and classify diseases at the morphological and molecular level, such as breast cancer,[5] and can foster the effective treatment of diseases such as in the case of a stroke.[6] AI methods are also increasingly involved in the evaluation of medical interventions, such as in the assessment of surgical performance.[7] Additionally, physicians increasingly face comparison with AI-based systems in terms of successful application of their expertise.[8]

With life-sciences research increasingly becoming part of medical treatment through the rapid translation of its findings into healthcare and through technology transfer, issues around the application of AI-based methods and products are becoming pertinent in medical care. AI applications, already ubiquitous, will only continue to multiply, permanently altering the healthcare system and in particular the individual doctor–patient relationship. Precisely because medical treatment has a direct impact on the life and physical integrity as well as the right of self-determination of patients involved, standards must be developed for the use of AI in healthcare. These guidelines are needed at the international level in order to ease the inevitable cross-border use of AI-based systems while boosting their beneficial impact on patients' healthcare. This would not only promote patient welfare and general confidence[9] in the benefits of medical AI, but would also help, for example, with the international marketing and uniform certification of AI-based medical devices,[10] thereby promoting innovation and facilitating trade.

A look at current statements, recommendations, and declarations by international organizations such as the United Nations Educational, Scientific and Cultural Organization (UNESCO), the World Health Organization (WHO), the Organisation for Economic Co-operation and Development (OECD), and the Council of Europe (CoE), as well as by non-governmental organizations such as the World Medical Association (WMA), shows that the importance of dealing with AI in as internationally uniform a manner as possible is already well recognized.[11] However, as will be shown in the following sections, international standardization

[5] A Binder and others, 'Morphological and Molecular Breast Cancer Profiling through Explainable Machine Learning' (*Nat Mach Intell*, 8 March 2021) www.nature.com/articles/s42256-021-00303-4.

[6] Medieninformation, 'Hirnschlag mit künstlicher Intelligenz wirksamer behandeln dank Verbundlernen' (*Universität Bern*, 9 March 2021). www.caim.unibe.ch/unibe/portal/fak_medizin/dept_zentren/inst_caim/content/e998130/e998135/e1054959/e1054962/210309_Medienmitteilung_InselGruppe_UniBE_ASAP_eng.pdf; WHO, *WHO Guideline: Recommendations on Digital Health Interventions for Health System Strengthening* (WHO/RHR/19.8, 2019) (hereafter WHO, *Recommendations on Digital Health*).

[7] JL Lavanchy and others, 'Automation of Surgical Skill Assessment Using a Three-Stage Machine Learning Algorithm' (2021) 11 *Sci Rep* 5197.

[8] M Nagendran and others, 'Artificial Intelligence versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies' (2020) *BMJ* 368:m689.

[9] See also European-Commission, 'High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI' (*European Commission*, 8 April 2019) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. The Guidelines of this group have benn subject to criticism, cf. M Veale, 'A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence' (2020) 11 European Journal of Risk Regulation 1, E1 doi:10.1017/err.2019.65.

[10] Cf., for example, FDA, 'Digital Health Software Precertification (Pre-Cert) Program' (*FDA*, 14 September 2020) www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program.

[11] CoE Commissioner for Human Rights, 'Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights' (*Council of Europe*, May 2019) 10 *et seq*. https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64; CoE Committee of Ministers, 'Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes' (1337th meeting of the Ministers' Deputies, Decl(13/02/2019)1, 13 February 2019) No. 9 https://search.coe.int/cm/pages/result_details.aspx?ObjectId=0900001680092dd4b; OECD, 'Recommendation of the Council on Artificial Intelligence' (OECD/LEGAL/0449, 22 November 2019) Section 2 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (SHS/BIO/PI/2021/1, 23 November 2021) II.7 https://unesdoc.unesco.org/ark:/48223/pf0000381137; WHO, 'Ethics and Governance for Artificial Intelligence for Health' (*WHO*, 21 June 2021) 2 *et seq*., 17 https://www.who.int/publications/i/item/9789240029200 (hereafter WHO, 'Ethics and Governance for Artificial Intelligence for Health'); and the following

for potential concrete AI applications in the various stages of medical treatment is not yet sufficient in terms of content. The situation is further complicated by the fact that the aforementioned instruments have varying degrees of binding force and legal effect. Following the identification of those gaps requiring regulation or guidance at the international level, the aim is to critically examine the international organizations and non-governmental organizations that could be considered for the job of closing them. In particular, when considering the spillover effect of the WMA's guidelines and statements on national medical professional law, it will be necessary to justify why the WMA is particularly suitable for creating regulations governing the scope of application of AI in the doctor–patient relationship.

II. APPLICATION AREAS OF AI IN MEDICINE ADDRESSED BY INTERNATIONAL GUIDELINES SO FAR

As sketched in the introduction, AI can be used to draw insights from large amounts of data at various stages of medical treatment. Thereby, AI can generally be defined as 'the theory and development of computer systems capable to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making [. . .]'.[12] Besides different types of AI systems as regards their autonomy and learning type,[13] a further distinction can be drawn in the context of decision-making in medical treatment as to whether AI is used as a decision aid or as a decision-maker.[14] While in the former case the physician retains the decision-making or interpretative authority over the findings of AI, in the latter case this does not normally apply, or only to a very limited extent. In any case, this distinction must be viewed critically insofar as even where AI is acting as a decision-maker the actors themselves, who are involved only to a small degree in the development and application of AI, each make individual decisions. Altogether, it is questionable whether decision-making can be assumed to be solely the result of AI's self-learning properties.[15] Given that AI can, at least potentially, be used in every

documents issued by the WMA: 'WMA Statement on Mobile Health' (66th WMA General Assembly, Russia, 20 February 2017) www.wma.net/policies-post/wma-statement-on-mobile-health/ (hereafter WMA, 'WMA Statement on Mobile Health'); 'WMA Statement on Augmented Intelligence in Medical Care' (70th WMA General Assembly, Georgia, 26 November 2019) www.wma.net/policies-post/wma-statement-on-augmented-intelligence-in-medical-care/ (hereafter WMA, 'WMA Statement on Augmented Intelligence'); 'WMA Statement on the Ethics of Telemedicine' (58th WMA General Assembly, Denmark, amended by 69th General Assembly, Iceland, 21 September 2020) No 1 www.wma.net/policies-post/wma-statement-on-the-ethics-of-telemedicine/ (hereafter WMA, 'WMA Statement on the Ethics of Telemedicine'); 'Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects' (18th WMA General Assembly, Finland, last amended by the 64th WMA General Assembly, Brazil, 9 July 2018) No 26 www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (hereafter Declaration of Helsinki).

[12] English Oxford Living Dictionary, 'Artificial Intelligence' www.lexico.com/definition/artificial_intelligence. In this chapter, when elaborating on AI methods, Deep Learning and Machine Learning are at the focus of considerations.

[13] Acatech, 'Machine Learning in der Medizintechnik' (*acatech*, 5 May 2020) 8, 11 www.acatech.de/publikation/machine-learning-in-der-medizintechnik/.

[14] Datenethikkommission, 'Gutachten der Datenethikkommission' (2020) 24, 28 (*Federal Ministry of the Interior, Building and Community*, 23 October 2019) www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6 (hereafter Datenethikkommission, 'Gutachten').

[15] The 'actorhood' of AI is discussed mainly from the perspectives of action theory and moral philosophy, which are not addressed in this chapter. Currently, however, it is assumed that AI-based systems cannot themselves be bearers of moral responsibility, because they do not fulfill certain prerequisites assumed for this purpose, such as freedom, higher-level intentionality, and the ability to act according to reason. On the abilities required for ethical machine reasoning and the programming features that enable them, cf. LM Pereira and A Saptawijaya, *Programming Machine Ethics* (2016). On the question of the extent to which AI-based systems can act, cf. C Misselhorn, *Grundfragen der Maschinenethik* (2018) and Chapter 3 in this volume. With regard to the legal assessment related to the 'actorhood' of AI systems and the idea of granting algorithmic systems with a high degree of autonomous legal personality in the

stage of medical treatment, from anamnesis to aftercare and documentation, and that the medical standards must be upheld, and that the patient must be kept informed at every stage, the gaps to be filled by an international guideline must be defined on the basis of a holistic view of medical treatment.

1. *Anamnesis and Diagnostic Findings*

The doctor–patient relationship usually begins with the patient contacting the doctor due to physical complaints, which the doctor tries to understand by means of anamnesis and diagnosis. Anamnesis includes the generation of potentially medically relevant information,[16] for example about previous illnesses, allergies, or regularly taken medications. The findings are collected by physical, chemical, or instrumental examinations or by functional testing of respiration, blood pressure, or circulation.[17]

An important AI application area is oncology. Based on clinical or dermatopathological images, AI can be used to diagnose and to classify skin cancer[18] or make a more accurate interpretation of mammograms for early detection of breast cancer.[19] Another study from November 2020 shows that AI could also someday be used to automatically segment the major organs and skeleton in less than a second, which helps in localizing cancer metastases.[20]

Among other things, wearables (miniaturized computers worn close to the body) and digital health applications[21] are also being developed for the field of oncology and are already being used by patients independently, for example, to determine their findings. For example, melanoma screening can be performed in advance of a skin cancer diagnosis using mobile applications such as store-and-forward teledermatology and automated smartphone apps.[22] Another 'use' case is monitoring patients with depression. The Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT) is seeking to complement existing apps for monitoring writing and reading behaviors of depressed patients with an app that provides AI-based speech analysis. The model recognizes speech style and word sequences and finds patterns indicative of depression. Using machine learning, it learns to detect depression in new patients.[23]

future ('electronic person'), the authors agree with the position of the German Data Ethics Commission, according to which this idea should not be pursued further. Cf. Datenethikkommission, 'Gutachten' (n 14) Executive Summary, 31 Nr 73. For this reason, the article only talks about AI *per se* for the sake of simplicity; this is neither intended to imply any kind of 'personalization' nor to represent a position in the debate about 'personalization' with normative consequences.

[16] A Laufs, BR Kern, and M Rehborn, '§ 50 Die Anamnese' in A Laufs, BR Kern, and M Rehborn (eds), *Handbuch des Arztrechts* (5th ed. 2019) para 1.

[17] C Katzenmeier, 'Arztfehler und Haftpflicht' in A Laufs, C Katzenmeier, and V Lipp (eds), *Arztrecht* (8th ed. 2021) para 4.

[18] TJ Brinker and others, 'Deep Learning Outperformed 136 of 157 Dermatologists in a Head-To-Head Dermoscopic Melanoma Image Classification Task' (2019) 113 *European Journal of Cancer* 47.

[19] A Esteva and others, 'Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks' (2017) 542 *Nature* 115.

[20] O Schoppe and others, 'Deep Learning-Enabled Multi-Organ Segmentation in Whole-Body Mouse Scans' (2020) 11 *Nat Commun* 5626.

[21] The Federal Institute for Drugs and Medical Devices keeps a record of all digital medical applications (*DiGA-Verzeichnis*) https://diga.bfarm.de/de/verzeichnis.

[22] S Chan and others, 'Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations' (2020) 10 Dermatol Ther *(Heidelb)* 365, 375.

[23] T Alhanai, M Ghassemi, and J Glass, 'Detecting Depression with Audio/Text Sequence Modelling of Interviews' (2018) *Proc Interspeech* 1716. Cf. also M Tasmin and E Stroulia, 'Detecting Depression from Voice' in *Canadian Conference on AI: Advances in Artificial Intelligence* (2019) 472.

As regards health apps and wearables, the WMA distinguishes between 'technologies used for lifestyle purposes and those which require the medical expertise of physicians and meet the definition of medical devices' and calls for the use of the latter to be appropriately regulated.[24] In its October 2019 statement, the WMA emphasizes that protecting the confidentiality and control of patient data is a core principle of the doctor–patient relationship.[25] In line with this, the CoE recommends that data protection principles be respected in the processing of health data, especially where health insurers are involved, and that patients should be able to decide whether their data will be disclosed.[26] The WHO draws attention to the complexity of the governance of data obtained from wearables, which may not have been collected initially for healthcare or research purposes.[27]

These statements provide a basic direction, but do not differentiate more closely between wearable technologies and digital health applications with regard to the type of use, the scope of health data collected and any transfer of this data to the physician. It is unclear how physicians should handle generated health data, such as whether they must conduct an independent review of the data or whether a plausibility check is sufficient to use the data when taking down a patient's medical history and making findings. The degree of transparency for the patient regarding the workings of the AI application as well as any data processing is also not specified. The implementation of a minimum standard or certification procedure could be considered here.

Telematics infrastructure can play a particularly important role at the beginning of the doctor-patient relationship. In its 2019 recommendations, the WHO distinguished between two categories of telemedicine. First, it recommends client-to-provider telemedicine, provided this does not replace personal contact between doctor and patient, but merely supplements it.[28] Here it agrees with the WMA's comprehensive 2018 statement on telemedicine, which made clear that telemedicine should only be used when timely face-to-face contact is not possible.[29] This also means that the physician treating by means of telemedicine should be the physician otherwise treating in person, if possible. This would require reliable identification mechanisms.[30] Furthermore, education, particularly about the operation of telemedicine, becomes highly important in this context so the patient can give informed consent.[31] The monitoring of patient safety, data protection, traceability, and accountability must all also be ensured.[32] After the first category of client-to-provider telemedicine has been established, the WHO also recommends provider-to-provider telemedicine as a second category, so that healthcare professionals, including physicians, can support each other in diagnoses, for example, by sharing images and video footage.[33] Thus, many factors must be clarified at the national level when creating a legal framework including licensing, cross-border telemedicine treatment, and use cases for remote consultations and their documentation.[34]

[24] WMA, 'WMA Statement on Mobile Health' (n 11).
[25] WMA, 'WMA Statement on Augmented Intelligence' (n 11).
[26] CoE Commissioner for Human Rights, 'Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights' (n 11) 10 *et seq*.
[27] WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 84.
[28] WHO, *Recommendations on Digital Health* (n 6).
[29] WMA, 'WMA Statement on the Ethics of Telemedicine' (n 11).
[30] Ibid, No 2.
[31] Ibid, No 4.
[32] WHO, *Recommendations on Digital Health* (n 6) 50.
[33] Ibid, 53 *et seq*.
[34] Ibid, 51.

In Germany, for example, the first regulations for the implementation of a telematics infrastructure have been in force since October 2020,[35] implementing the recommendations of the WHO and the WMA among others. There, the telematics infrastructure is to be an interoperable and compatible information, communication, and security infrastructure that serves to network service providers, payers, insured persons, and other players in the healthcare system and in rehabilitation and care.[36] This infrastructure is intended to enable telemedical procedures, for instance, for video consultation in SHI-accredited medical care.[37] For this purpose, § 365 SGB V explicitly refers to the high requirements of the physician's duty of disclosure for informed consent pursuant to § 630e BGB (German Civil Code)[38], which correspond to those of personal treatment.

Telemedicine should be increasingly used to close gaps in care and thus counteract disadvantages, especially in areas with a poorer infrastructure in terms of local medical care.[39] To this end, it could be helpful to identify for which illnesses telemedical treatment is sufficient, or determine whether such a treatment can already be carried out at the beginning of the doctor–patient relationship. One example thereof are the large-scale projects in the German region of Brandenburg, where, for example, patients' vital signs were transmitted telemedically as part of a study to provide care for heart patients.[40] In the follow-up study, AI is now also being used to prepare the vital data received at the telemedicine center for medical staff.[41]

2. *Diagnosis*

The findings must then be evaluated professionally, incorporating ideas about the causes and origins of the disease, and assigned to a clinical picture.[42]

Accordingly, AI transparency and explicability become especially important in the area of diagnosis. In its October 2019 statement, the WMA pointed out that physicians need to understand AI methods and systems so that they can make medical recommendations based on them, or refrain from doing so if individual patient data differs from the training data used.[43] It can be concluded, just as UNESCO's Ad Hoc Expert Group (AHEG) directly stated in its September 2020 draft, that AI can be used as a decision support tool, but should not be used as a

[35] Law on the Protection of Electronic Patient Data within the Telematic Infrastructure (Gesetz zum Schutz elektronischer Patientendaten in der Telematikinfrastruktur), BGBl. 2020, 2115.

[36] Social Security Statute Book V – Statutory Health Insurance (SGB V), Article 1 of the Act of 20 December 1988 (Federal Law Gazette [*Bundesgesetzblatt*] I page 2477, 2482), last amended by Artikel 1b of the Act of 23 Mai 2022 (Federal Law Gazette I page 760), §306(1) sentence 2.

[37] § 364 *et seq*. SGB V.

[38] Civil Code in the version promulgated on 2 January 2002 (Federal Law Gazette [*Bundesgesetzblatt*] I page 42, 2909; 2003 I page 738), last amended by Article 2 of the Act of 21 December 2021 (Federal Law Gazette I page 5252).

[39] These advantages, which also increase the acceptance of health workers for digital health interventions, are described by the WHO: World Health Organization, *Recommendations on Digital Health* (n 6) 34. In addition, the WHO has recently suggested exploring whether the introduction and use of AI in healthcare exacerbates the digital divide. Ultimately, AI using telemedicine should reduce the gap in access to healthcare and ensure equitable access to quality care, regardless of geographic and other demographic factors: WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 74.

[40] For more information cf. Charité, 'Fontane' https://telemedizin.charite.de/forschung/fontane/.

[41] For more information cf. Charité, 'Telemed5000' https://telemedizin.charite.de/forschung/telemed5000/.

[42] A Laufs, BR Kern, and M Rehborn, '§ 52 Die Diagnosestellung' in A Laufs, BR Kern, and M Rehborn (eds), *Handbuch des Arztrechts* (5th ed. 2019) para 7 *et seq*.

[43] WMA, 'WMA Statement on Augmented Intelligence' (n 11).

decision-maker replacing human decision and responsibility.[44] The WHO also recommends the use of AI as a decision support tool only when its use falls within the scope of the physicians' current field of work, so that the physicians provide only the services for which they have been trained.[45]

There is no clarification as to the extent to which transparency is required of the physician as regards AI algorithms and decision logic. A distinction should be made here between open-loop and closed-loop systems.[46] An open-loop system, in which the output has no influence on the control effect of the system, is generally easier to understand and explain, allowing stricter requirements to be placed on the control of AI decisions and treatments based on them. On the other hand, it is more difficult to deal with closed-loop systems in which the output depends on the input because the system has one or more feedback loops between its output and input. In addition, there is the psychological danger that the physician, knowing the nature of the system and its performance, may consciously or unconsciously exercise less rigorous control over the AI decision. It is, therefore, necessary to differentiate between both the type of system and the use of AI dependent on its influence as a decision aid in order to identify the degree of necessary control density, from simple plausibility checks to more intensive review obligations of the physician. It is also clear that there is a need to explain which training data and patient data were processed and influenced the specific diagnosis and why other diagnoses were excluded.[47] This is particularly relevant in the area of personalized and stratified diagnostics. In this context, the previously rejected possibility of AI as a decision-maker and the physician's ultimate decision-making authority could be re-explored and enabled under specific, narrowly defined conditions depending on the type of application and the type and stage of the disease, which could reduce the burden on healthcare infrastructure.

3. *Information, Education, and Consent*

Before treatment in accordance with the diagnosis can be started, the patient must be provided with treatment information to ensure that the patient's behavior is in line with the treatment and with economic information on the assumption of costs by the health insurance company.[48] In addition, information about the diagnosis, risks, and course of treatment as well as real alternatives to treatment is a prerequisite for effective patient consent.[49]

The WMA's Declaration of Helsinki states that information and consent should be obtained by a person qualified to give treatment.[50] The CoE's May 2019 paper also requires that the user or patient be informed when AI is used to interact with them in the context of treatment.[51] It is

[44] Ad Hoc Expert Group (AHEG) for the preparation of a draft text of a recommendation on ethics of artificial intelligence, 'Outcome Document: First Draft of the Recommendation on the Ethics of Artificial Intelligence' (September 2020) No 36 https://unesdoc.unesco.org/ark:/48223/pf0000373434.

[45] WHO, 'Guideline: Recommendations on Digital Health Interventions for Health System Strengthening' (n 6) 65; WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) p 6.

[46] Acatech, 'Machine Learning in der Medizintechnik' (n 13) 11.

[47] WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 106 *et seq*.

[48] C Katzenmeier, 'Aufklärungspflicht und Einwilligung' in A Laufs, C Katzenmeier, and V Lipp (eds), *Arztrecht* (8th ed. 2021) para 16, 21.

[49] Ibid, para 14.

[50] WMA, 'Declaration of Helsinki' (n 11) No 26.

[51] CoE Commissioner for Human Rights, 'Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights' (n 11) 10 *et seq*.

questionable whether a general duty of disclosure can be derived from this for every case in which AI is involved in patient care, even if only to a very small extent. The WHO's recent guidelines emphasize the increasing infeasibility of true informed consent particularly for the purpose of securing privacy.[52] In any case, there is currently a lack of guidance regarding the scope of the duty to disclose the functioning of the specific AI.

It would also be conceivable to use AI to provide information itself, for instance, through a type of chatbot system, if it had the training level and the knowledge of a corresponding specialist physician and queries to the treating physician remained possible. In any case, if this is rejected with regard to the physician's ultimate decision-making authority, obtaining consent with the help of an AI application after information has been provided by a physician could be considered for time-efficiency reasons.

4. *Treatment and Aftercare*

Treatment is selected based on a diagnosis, the weighing of various measures and risks, the purpose of the treatment and the prospects of success according to its medical indication. After treatment is complete, monitoring, follow-up examinations, and any necessary rehabilitation take place.[53]

According to the Declaration of Helsinki, the physician's reservation and compliance with medical standards both apply, particularly in the therapeutic treatment of the patient.[54] No specific regulation has been formulated to govern the conditions under which AI used by physicians in treatment fulfil medical standards, and it is not clear whether it is necessary for them to meet those standards at all or whether even higher requirements should be placed on AI.[55] In addition, the limitations on a physician's right to refuse the use of AI for treatment are unclear. It is possible that the weight of the physician's ultimate decision-making authority could be graded to correspond to the measure and the risks of the treatment, especially in the context of personalized and stratified medicine, so that, depending on the degree of this grading, treatment by AI could be made possible.

AI allows the remote monitoring of health status via telemedicine, wearables, and health applications, for example, by monitoring sleep rhythms, movement profiles, and dietary patterns, as well as reminders to take medication. This is of great advantage especially in areas with poorer healthcare structures.[56] For example, a hybrid closed-loop system for follow-up care has already been developed for monitoring diabetes patients that uses AI to automate and personalize diabetes management. The self-learning insulin delivery system autonomously monitors the user's insulin level and delivers an appropriate amount of insulin when needed.[57] Furthermore, a December 2020 study shows that AI can also be used in follow-up and preventive care for young patients who have suffered from depression or have high-risk syndromes to predict the

[52] WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 40 *et seq*., with some suggestions in Box 4, 48, 82, and 90.

[53] C Katzenmeier, 'Arztfehler und Haftpflicht' in A Laufs, C Katzenmeier, V Lipp (eds), *Arztrecht* (8th ed. 2021) para 4.

[54] WMA, 'Declaration of Helsinki' (n 11) No 12, No 10.

[55] WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 77.

[56] See Sub-section II 1.

[57] For further information see C Amadou, S Franc, PY Benhamou, S Lablanche, E Huneker, G Charpentier, A Penfornis & Diabeloop Consortium 'Diabeloop DBLG1 Closed-Loop System Enables Patients With Type 1 Diabetes to Significantly Improve Their Glycemic Control in Real-Life Situations Without Serious Adverse Events: 6-Month Follow-up' (2021) 44 *Diabetes care* 3, 844.

transition to psychosis in a personalized way.[58] Meanwhile, follow-up also includes monitoring or digital tracking using an electronic patient file or other type of electronic health record so that, for example, timely follow-up examinations can be recommended. This falls under the digital tracking of clients' health status and services, which the WHO recommends in combination with decision support and targeted client communication (if the existing healthcare system can support implementation and the area of application falls within the area of competence of the responsible physician and the protection of patient data is ensured).[59]

However, there is as of yet no regulatory framework for the independent monitoring and initiation of AI-measures included in such applications. Apart from the need for regulation of wearables and health applications,[60] there is also a need for regulation of the transmission of patient data to AI, which must be solved in a way that is compliant with data protection rules.[61]

5. *Documentation and Issuing of Certificates*

The course of medical treatment is subject to mandatory documentation.[62] There is no clarification as to what must be documented and the extent of documentation required in relation to the use of AI in medical treatment. A documentation obligation could, for example, extend to the training status of AI, any training data used, the nature of its application, and its influence on the success of the treatment.

Both economically and in terms of saving time, it could make sense to employ AI at the documentation stage in addition to its use during treatment, as well as for issuing health certificates and attestations, leaving more time for the physician to interact with the patient.

6. *Data Protection*

The use of AI in the medical field must also be balanced against the data protection law applicable in the respective jurisdiction. In the EU this would be the General Data Protection Regulation (GDPR)[63] and the corresponding member state implementation thereof.

The autonomy[64] and interconnectedness[65] of AI alone pose data protection law challenges, and these are only exacerbated when AI is used in the context of medical treatment due to the sensitivity of personal health-related data. For example, as Article 22(1) of the GDPR protects

[58] N Koutsouleris and others, 'Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients with Clinical High-Risk Syndromes and Recent-Onset Depression' (*JAMA Psychiatry*, 2 December 2020) https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2773732.

[59] WHO, 'Guideline: Recommendations on Digital Health Interventions for Health System Strengthening' (n 6) 69 *et seq*. Considering the use of AI to extend 'clinical' care beyond the formal health-care system based on monitoring: WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 9 *et seq*.

[60] See Sub-section II 1.

[61] Cf. Sub-section II 6.

[62] For example, in civil law provisions in Germany according to § 63of BGB and for research studies based on the international standards of the WMA according to the Declaration of Helsinki (n 11) No 22.

[63] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88.

[64] R Konertz and R Schönhof, *Das technische Phänomen „Künstliche Intelligenz" im allgemeinen Zivilrecht. Eine kritische Betrachtung im Lichte von Autonomie, Determinismus und Vorhersehbarkeit* (2020) 69.

[65] H Zech, 'Künstliche Intelligenz und Haftungsfragen' (2019) *ZfPW*, 118, 202.

data subjects from adverse decisions based solely on automated processing, at least the final decision must remain in human hands.[66]

The processing of sensitive personal data such as health data is lawful if the data subject has given his or her express consent.[67] Effective consent is defined as words or actions given voluntarily and with knowledge of the specific factual situation.[68] A person must, therefore, know what is to happen to the data. In order to consent to treatment involving the use of AI, the patient would have to be informed accordingly.[69] However, it is difficult to determine how to inform the patient about the processing of the data if the data processing procedure changes autonomously due to the self-learning property of the AI. Broad consent[70] on the part of the patient is challenging as they would be consenting to unforeseeable developments and would consequently have precisely zero knowledge of the specific factual situation at the time of consent, effectively waiving the exercise of part of their right to self-determination. The GDPR operationalizes the fundamental right to the protection of personal data by defining subjective rights of data subjects, but it is questionable to what extent these rights would enable the patient to intercept and control data processing. The role of the patient, on the other hand, would be strengthened by means of dynamic information and consent[71], as the patient could give his or her consent bit by bit over the course of treatment using AI. The challenge here would be primarily on the technical side, as an appropriate organization and communication structure would have to be created to inform the patient about further, new data processing by the AI.[72] The patient would have to be provided with extensive information not only about the processed data but also about the resulting metadata if the latter reveals personally identifiable information, not least in order to revoke their consent, if necessary, in a differentiated way,[73] and to arrange for the deletion of their data.

Correspondingly, Articles 13 and 14 of the GDPR provide for information obligations and Article 17 of the GDPR for a right to deletion. A particular problem here is that the patient data fed in becomes the basis for the independent development of the AI and can no longer be deleted. Technical procedures for anonymizing the data could in principle help here, although this would be futile in a highly contextualized environment.[74] The use of different pseudonymization types (for instance noise) to lower the chance of re-identifiability might also be worth considering. This might, however, render the data less usable.[75] In any case, the balancing of the

[66] B Buchner, 'DS-GVO Art. 22' in J Kühling and B Buchner (eds), *Datenschutzgrundverordnung BDSG Kommentar* (3rd ed. 2020) para 14 *et seq.* P Schantz and HA Wolff, *Das neue Datenschutzrecht* (2017) recital 736.

[67] GDPR, Article 9(2)(a), in conjunction with Article 6(1)(a) GDPR or Article 6(1)(b) GDPR (doctor–patient relationship as a contractual obligation under civil law).

[68] D Kampert, 'DSGVO Art. 9' in G Sydow (ed), *Europäische Datenschutzgrundverordnung* (2nd ed. 2018) para 14.

[69] For challenges see Sub-section II 3.

[70] GDPR, Recital 33. WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 84 *et seq.*

[71] Cf. instead of many: HC Stoeklé and others, 'Vers un consentement éclairé dynamique' [Toward Dynamic Informed Consent] (2017) 33 *Med Sci* (Paris) 188. I Budin-Ljøsne and others, 'Dynamic Consent: A Potential Solution to Some of the Challenges of Modern Biomedical Research' (2017) 18(1) *BMC Med Ethics* 4. WHO, 'Ethics and Governance for Artificial Intelligence for Health' (n 11) 82.

[72] Information obligations in the course of broad and dynamic consent: Datenschutzkonferenz, 'Beschluss der 97. Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zu Auslegung des Begriffs "bestimmte Bereiche wissenschaftlicher Forschung" im Erwägungsgrund 33 der DS-GVO' (3 April 2019) www.datenschutzkonferenz-online.de/media/dskb/20190405_auslegung_bestimmte_bereiche_wiss_forschung.pdf.

[73] For problems with this, see Sub-section II 3.

[74] PHG Foundation, 'The GDPR and Genomic Data: The Impact of the GDPR and DPA 2018 on Genomic Healthcare and Research' (2020) 44 *et seq.* www.phgfoundation.org/media/123/download/gdpr-and-genomic-data-report.pdf?v=1&inline=1.

[75] Ibid, 167.

conflicting legal positions could lead to a restriction of deletion rights.[76] This in turn raises the question of the extent to which consent, which may also be dynamic, could be used as a basis of legitimacy for the corresponding processing of the data, even after the appropriate information about the limitations has been provided. In order to avoid a revocation of consent leading to the exclusion of certain data, other legal bases for processing are often proposed.[77] This often fails to take into account that erasure rights and the right to be forgotten may lead to a severe restriction of processing regardless. Additionally, the compliance with other rights, such as the right to data portability,[78] might be hampered or limited due to the self-learning capabilities of AI, with the enforcement of such rights leading to the availability of a given data set or at least its particular patterns throughout different applications, obstructing the provision of privacy through control over personal data by the data subject.

Because of the AI methods involved in processing patients' sensitive data and its regularly high contextualization, the likeliness of anonymized data, or data thought to be anonymized, becoming re-identifiable is also higher. Based on AI methods of pattern recognition, particular combinations of data fed into a self-learning AI system might be re-identified if the AI system trained with that data is later, in the course of its application, confronted with the same pattern. In this way, even if data or data sets were originally anonymized before being fed into an AI system, privacy issues may emerge due to the high contextuality of AI applications and their self-learning characteristics.[79] As a consequence, privacy issues will not only be relevant when data is moved between different data protection regimes, but also when data is analysed. However, the fact of re-identifiability might remain hidden for a considerable time.

Once re-identifiability is discovered, the processing of affected personal data will fall within the scope of application of the GDPR. Although at first glance this implies higher protection, unique characteristics of AI applications pose challenges to safeguard the rights of data subjects. A prominent example hereby is the right to be forgotten. As related to informational self-determination, the right to be forgotten is intended to prevent the representation and permanent presence of information in order to guarantee the possibility of free development of the personality. With the right to be forgotten, the digital, unlimited remembrance and retrieval of information is confronted with a claim to deletion in the form of non-traceability.[80] The concept of forgetting does not necessarily include a third party but does imply the disappearance of information as such.[81] Relating to data fed into AI applications, the connection between one's own state of ignorance and that of others, as well as their forgetting, including AI's ability to forget, remains decisive. Even if the person is initially able to ward off knowledge, it is still conceivable that others might experience or use this knowledge (relying on the increased re-identifiability of the data), and then in some form, even if derivatively, connect the data back to the individual. In this respect, forgetting by third parties is also relevant as an upstream protection for one's own forgetting. Furthermore, the right to be forgotten becomes an indispensable condition for many further rights of the person concerned. Foreign and personal forgetting are necessary, if information processing detaches itself from the person concerned and

[76] Cf. GDPR, Article 17(3)(d). T Herbst, 'DS-GVO Art. 22' in J Kühling, B Buchner (eds), *Datenschutzgrundverordnung BDSG Kommentar* (3rd ed. 2020) para. 81 *et seq*.

[77] For special categories of personal data, cf. the exemptions defined in Article 9(2) GDPR.

[78] Cf. GDPR, Article 20.

[79] Cf. instead of many others: B Murdoch, 'Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era' (2021) 22(1) *BMC Medical Ethics* 122.

[80] CJEU, Case C-131/12 *Google Spain v Gonzalez* [2014] ECLI:EU:C:2014:317 para 87 *et seq*.

[81] OJ Gstrein, *Das Recht auf Vergessenwerden als Menschenrecht* (2016) 111.

becomes independent to then be fed back into their inherently internal decision-making processes, leveraging the realization of (negative) informational self-determination.[82]

7. Interim Conclusion

The variety of already-existing uses of AI in the context of medical treatment, from initial contact to follow-up and documentation, shows the increasingly urgent need for uniform international standards, not least from a medical ethics perspective. Above all, international organizations such as the WHO and non-governmental organizations such as the WMA have set an initial direction with their statements and recommendations regarding the digitization of healthcare. However, it is striking that, on the one hand there are more recent differentiated recommendations for the application of AI in medical treatment in general but these are not directed at physicians in particular and that, on the other hand, the entire focus of such recommendations is regularly on individual subareas and on the governance in healthcare without a comprehensive examination of possible applications in the physician–patient relationship. Medical professionals, especially physicians, are thus exposed to different individual and general recommendations in addition to the technical challenges already posed by AI. This could lead to uncertainties and differing approaches among physicians and could ultimately have a chilling effect on innovation. Guidelines from a competent international organization or professional association that cover the use of AI in all stages of medical treatment, especially from the physician's perspective, would therefore be desirable.

III. INTERNATIONAL GUIDANCE FOR THE FIELD OF AI APPLICATION DURING MEDICAL TREATMENT

1. International Organizations and Their Soft-Law Guidance

Both the WHO and the UNESCO are specialized agencies of the United Nations traditionally responsible for the governance of public health.[83] The WHO has been regularly engaged in fieldwork as an aid to research ethics committees, but has recently increasingly moved into developing guidance within the area of public health and emerging technologies.[84] UNESCO derives its responsibility for addressing biomedical issues from the preamble to its statutes and, at the latest since the 2005 Bioethics Declaration,[85] has indicated that it intends to assume the role of international coordinator in the governance of biomedical issues.[86] Here, UNESCO relies on an institutionalization of its ethical mandate in the form of the International Bioethics Committee.[87] Currently, both organizations' key activity in this area focuses on setting standards: since the development of science and technology has become increasingly global in order to

[82] For this in-depth analysis of the right to be forgotten, cf. F Molnár-Gábor, 'Das Recht auf Nichtwissen. Fragen der Verrechtlichung im Kontext von Big Data in der modernen Biomedizin' in G Duttge and Ch Lemke (eds), *Das sogenannte Recht auf Nichtwissen. Normatives Fundament und anwendungspraktische Geltungskraft* (2019) 83, 99 *et seq.*

[83] JE Alvarez, *International Organizations as Law-Makers* (2005) 4, 6 *et seq*. Due to the regional character of the Council of Europe, its instruments are not further elaborated on here.

[84] WHO, 'Global Health Ethics' https://apps.who.int/iris/handle/10665/164576.

[85] UNESCO, 'Universal Declaration on Bioethics and Human Rights, 19 October 2005, Records of the UNESCO General Conference, 33rd Session, Paris, 3–21 October 2005' (33 C/Resolution 36) 74 *et seq.*

[86] Constitution of the UNESCO, 4 UNTS 275, UN Reg No I-52 (hereafter UNESCO-Constitution).

[87] F Molnar-Gabor, *Die internationale Steuerung der Biotechnologie am Beispiel neuer genetischer Analysen* (2017) 202 *et seq.*

accompany progress, provide the necessary overview, and ensure equal access to the benefits of scientific development, there is a need for global principles in various areas that member states can apply as a reference framework for establishing specific regulatory measures.

Such global principles are developed by both organizations, notably in the form of international soft law.[88] According to prevailing opinion, this term covers rules of conduct of an abstract, general nature that have been enacted by subjects of international law but which cannot be assigned to any formal source of law and are not directly binding.[89] However, soft law instruments cannot be reduced to mere political recommendations but can unfold *de facto* 'extra-legal binding effect', despite their lack of direct legal binding force.[90] International soft law can also be used as an indicator of legal convictions for the interpretation of traditional sources of international law such as treaties.[91] Furthermore, it can provide evidence of the emergence of customary law and lead to obligations of good faith.[92] Soft law can also serve the further development of international law: It can often be a practical aid to consensus-building and can also provide a basis for the subsequent development of legally binding norms.[93] Such instruments can also have an effect on national legal systems if, for example, they are introduced into national legal frameworks through references in court decisions.[94]

Criticism of UNESCO's soft law documents is mainly directed at the participation in and deliberation of decisions.[95] Article 3(2) of the Statutes of the International Bioethics Committee of UNESCO (IBC Statutes)[96] prescribes the nomination of eminent experts to the member states.[97] Although the IBC's reports generally show a particular sensitivity to normative challenges of emerging health technologies, the statute allows the involvement of external experts in the drafting processes – an option that has not been widely used by the IBC in the course of preparing the main UNESCO declaration in the area of bioethics.[98] The IBC's reports are regularly revised and finalized by the Inter-Governmental Bioethics Committee (IGBC), which represents the member states' governments.[99] This is justified by the fact that the addressees and primary actors in the promotion and implementation of the declarations are

[88] On the advantages of international soft law compared to international treaties when it comes to the regulation of biomedicine cf. A Boyle, 'Some Reflections on the Relationship of Treaties and Soft Law' (1999) 48 *International and Comparative Law Quarterly* 901, 902 *et seq*., 912 *et seq*.; R Andorno, *Principles of International Biolaw. Seeking Common Ground at the Intersection of Bioethics and Human Rights* (2013) 39 *et seq*.; W Höfling, 'Professionelle Standards und Gesetz' in HH Trute and others (eds), *Allgemeines Verwaltungsrecht – zur Tragfähigkeit eines Konzepts, Festschrift für Schmidt-Aßmann zum 70. Geburtstag* (2008) 45, 52.

[89] M Bothe, 'Legal and Non-Legal Norms: A Meaningful Distinction in International Relations?' (1980) 11 *Netherlands Yearbook of International Law* 65, 67 *et seq*.

[90] J Klabbers, *An Introduction to International Institutional Law* (2nd ed. 2009) 183.

[91] H Hilgenberg, 'Soft Law im Völkerrecht' (1998) 1 *Zeitschrift für Europarechtliche Studien* 81, 100 *et seq*.

[92] M Goldmann, *Internationale öffentliche Gewalt* (2015) 34, 60 *et seq*., 187 *et seq*., 199 *et seq*.

[93] I Venzke, *How Interpretation Makes International Law* (2012) 380.

[94] TA Faunce, 'Will International Human Rights Subsume Medical Ethics? Intersections in the UNESCO Universal Bioethics Declaration' (2005) 31 *Journal of Medical Ethics* 173, 176; D Thürer, 'Soft Law' in R Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (2009) recital 2.

[95] Cf. instead of many others: A Langlois, *Negotiating Bioethics* (2013) (hereafter Langlois, *Negotiating Bioethics*) 144.

[96] Statutes of the International Bioethics Committee of UNESCO (IBC), Adopted by the Executive Board at its 154th Session, on 7 May 1998 (154 EX/Dec. 8).

[97] F Molnár-Gábor, *Die internationale Steuerung der Biotechnologie am Beispiel neuer genetischer Analysen* (2017) 298 *et seq*.

[98] Ibid, 301.

[99] Statutes of the International Bioethics Committee of UNESCO (IBC) (n 96) Article 11. Cf. Rules of Procedure of the Intergovernmental Bioethics Committee (IGBC), Adopted by IGBC at its 3rd session on 23 June 2003 in Paris and amended at its 5th session on 20 July 2007 and at its 7th session on 5 September 2011 (SHS/EST/IGBC-5/07/CONF.204/7 Rev) Article 1.

the member states.[100] However, only 36 member states are represented on the committee at once, which is just one-fifth of all UNESCO member states. Moreover, the available seats do not correspond to the number of member states in each geographic region. While approximately every fourth member state is represented from Western Europe and the North American states, only approximately every fifth member state is represented from the remaining regions.[101]

2. *The World Medical Association*

The highest ethical demands are to be made of physicians within the scope of their professional practice because of their great responsibility towards the life, the bodily integrity and the right of self-determination of the patient.[102] In order to establish such an approach worldwide, the WMA was founded in 1947 following the Nuremberg trials as a reaction to the atrocities of German physicians in the Third Reich.[103] Today, as a federation of 115 national medical associations, it promotes 'the highest possible standards of medical ethics' and 'provides ethical guidance to physicians through its Declarations, Resolutions and Statements'.[104] Unlike the international organizations described earlier, it is not a subject of international law, but a non-governmental organization that acts autonomously on a private law basis. As it is not based on a treaty under international law, the treaties it concludes with states would not be subject to international treaty law either.[105] The WMA is, therefore, to be treated as a subject of private law.

Such subjects of private law are well able to focus on specific topics to provide guidance and are, therefore, in a good position to address the challenges of biomedical issues. However, the Declaration of Helsinki and other declarations of the WMA have no legally binding character as resolutions of an international alliance of national associations under private law and can only be regarded as a codification of professional law, not as international soft law.[106] Yet, as will also be shown using the example of Germany, they are well integrated into national professional laws.

One criticism of the WMA's decision-making legitimacy is that its internal deliberation is not very transparent and takes place primarily within the Council and the relevant committee(s), whose members are designated by the Council from its own members.[107] This means that some national medical associations barely participate in the deliberation. Currently, for example, only nine out of 27 Council members are from the Asian continent and one out of 27 from the

[100] Critically on this Langlois, *Negotiating Bioethics* (n 95) 56.

[101] F Molnár-Gábor, *Die internationale Steuerung der Biotechnologie am Beispiel neuer genetischer Analysen* (2017) 299 *et seq*. For the critical assessment of the Inter-Governmental Meeting of Experts, cf. Langlois, *Negotiating Bioethics* (n 95) 56. The distribution of seats and the election take place according to the decision of the Executive Council: 155 EX/Decision 9.2, Paris, 03.12.1998. According to this, Group I (Western Europe and the North American States) has seven seats, Group II (Eastern Europe) has four, Group III (Latin America and the Caribbean States) has six, Group IV (Asia and the Pacific States) has seven, and Group V (Africa [eight] and the Arab States [four]) has a total of twelve seats.

[102] W Spann, 'Ärztliche Rechts- und Standeskunde' in A Ponsold (ed), *Lehrbuch der Gerichtlichen Medizin* (1957) 4.

[103] T Richards, 'The World Medical Association: Can Hope Triumph Over Experience?' (1994) *BMJ*, 308 (hereafter Richards, 'The World Medical Association').

[104] See official homepage: WMA, 'About Us' www.wma.net/who-we-are/about-us/ (hereafter WMA, 'About Us').

[105] S Vöneky, 'Rechtsfragen der Totalsequenzierung des menschlichen Genoms in internationaler und nationaler Perspektive' (2012) *Freiburger Informationspapiere zum Völkerrecht und Öffentlichen Recht* 4, note 16, https://www.jura.uni-freiburg.de/de/institute/ioeffr2/downloads/online-papers/fip_4_2012_totalsequenzierung.pdf.

[106] Ibid.

[107] On the decision-making process M Chang, 'Bioethics and Human Rights: The Legitimacy of Authoritative Ethical Guidelines Governing International Clinical Trials' in S Voeneky and others (eds), *Ethics and Law: The Ethicalization of Law* (2013) 177, 210 (hereafter Chang, 'Bioethics and Human Rights').

African continent,[108] which is disproportionate compared to their population densities. Council bills are debated and discussed in the General Assembly but, given the lack of time and number of bills to be discussed, the Assembly does not have as much influence on the content as the Council and Committees.[109] Each national medical association may send one voting delegate to the General Assembly. In addition, they may send one additional voting member for every ten thousand members for whom all membership dues have been paid.[110] This makes the influence of a national medical association dependent, among other things, on its financial situation. Of additional concern is the fact that these national medical associations do not necessarily represent all types of physicians, because membership is not mandatory in most countries.[111] Moreover, other professional groups affected by the decisions of the WMA are not automatically heard.[112] As a consequence of the WMA's genesis as a result of human experimentation by physicians in the Third Reich and the organization's basis in the original Declaration of Helsinki,[113] the guidelines of the WMA are based primarily on American- or European-influenced medical ethics, although the membership of the WMA is more diverse.[114]

3. *Effect of International Measures in National Law*

a. Soft Law

Declarations of UNESCO as international soft law[115] are adopted by the General Conference.[116] They cannot be made binding on the member states and are not subject to ratification. They set forth universal principles to which member states 'wish to attribute the greatest possible authority and to afford the broadest possible support'.[117] Additionally, UNESCO's Constitution does not include declarations among the proposals which may be submitted to the General Conference for adoption[118], although the General Conference can, in practice, adopt a document submitted to it in the form of a declaration.[119] Besides their contribution to shaping and developing binding norms and helping the interpretation of international law, soft law norms may also have immediate legal effects in the field of good faith, even if this does not change the non-legal nature of soft law.[120] This effect has particular relevance in the field of medicine and bioethics.

[108] See official homepage: WMA, 'About Us' (n 104).
[109] Chang, 'Bioethics and Human Rights' (n 107), 177, 209. Cf. Richards, 'The World Medical Association' (n 103).
[110] Chang, 'Bioethics and Human Rights' (n 107), 177, 209 *et seq*. The threshold was 50,000 members a few years ago. Cf. Richards, 'The World Medical Association' (n 103).
[111] Chang, 'Bioethics and Human Rights' (n 107), 177, 214.
[112] Cf. Chang, 'Bioethics and Human Rights' (n 107), 177, 212.
[113] WMA, 'Declaration of Helsinki' (n 11).
[114] This medical ethics has been condensed into the four bioethical principles of autonomy, beneficence, non-maleficence, and justice (as set down by Beauchamp and Childress). TL Beauchamp and JF Childress, *Principles of Biomedical Ethics* (8th ed. 2012). For criticism on principalism cf. U Wiesing, 'Vom Nutzen und Nachteil der Prinzipienethik für die Medizin' in O Rauprich and F Steger (eds), *Prinzipienethik in der Biomedizin. Moralphilosophie und medizinische Praxis* (2005) 74, 77 *et seq*.
[115] S Voeneky, *Recht, Moral und Ethik* (2010) 383.
[116] UNESCO, 'Declarations' http://portal.unesco.org/en/ev.php-URL_ID=12027&URL_DO=DO_TOPIC&URL_SECTION=-471.html.
[117] UNESCO, 'General Introduction to the Standard-Setting Instruments of UNESCO' http://portal.unesco.org/en/ev.php-URL_ID=23772&URL_DO=DO_TOPIC&URL_SECTION=201.html (hereafter UNESCO, 'General Introduction').
[118] Article 4(4) UNESCO-Constitution (n 86).
[119] UNESCO, 'General Introduction' (n 117).
[120] D Thürer, 'Soft Law' in R Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (2009) recital 27 (hereafter Thürer, 'Soft Law').

The principle of good faith requires relevant actors not to contradict their own conduct.[121] Accordingly in the area of soft law, it legally protects expectations produced by these norms insofar as it is justified by the conduct of the parties concerned.[122] UNESCO itself states that declarations may be considered to engender a strong expectation that members states will abide by them on the part of the body adopting them. Consequently, insofar as the expectation is gradually justified by state practice, a declaration may by custom become recognized as laying down rules that are binding upon states.[123]

b. Incorporation of WMA Measures into Professional Law

At the national level, professional law has an outstanding importance for physicians. In Germany, for example, the definition of individual professional duties is the responsibility of the respective state medical association, which issues professional regulations in the form of statutes. The autonomy of the statutes is granted to the state medical associations by virtue of state law and is an expression of the functional self-administration of the medical associations. In addition to defining professional duties, the state medical associations are also responsible for monitoring physicians' compliance with these duties.[124] Due to the compulsory membership of physicians in the state medical associations, the professional law or respective professional code of conduct is obligatory for each individual physician.[125] The state medical associations are guided in terms of content by the Model Code of Professional Conduct for Physicians (MBO-Ä),[126] which is set out by the German Medical Association (*Bundesärztekammer*) as the association of state medical associations (and thus the German member of the WMA). If a declaration or statement is adopted at the international level by the WMA, the German Medical Association will incorporate the contents into the MBO-Ä, not least if it was involved in the deliberation. In addition to the statutes issued by the state medical associations, regulations on the professional conduct of physicians are found partly in federal laws such as the Criminal Code,[127] or the Civil Code,[128] and partly in state laws such as hospital laws. Regardless of which regulations are applicable in a specific case, the physician must always carry out the treatment of a patient in accordance with medical standards.[129]

The medical standard to be applied in a specific case must be interpreted according to the circumstances of the individual case, taking into account what has objectively emerged as medical practice in scientific debate and practical experience and is recognized in professional circles as the path to therapeutic success, as well as what may be expected subjectively from the respective physician on average.[130] Any scientific debate about the application of AI in medical treatment on the level of the WMA would take place in professional circles and could thereby

[121] M Kotzur, 'Good Faith (Bona Fide)' in R Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (2009) recital 25.
[122] Thürer, 'Soft Law' (n 120) recital 27. Cf. definition by M Goldmann, *Internationale öffentliche Gewalt* (2015) p. 3.
[123] UNESCO, 'General Introduction' (n 117).
[124] Compare V Lipp, 'Ärztliches Berufsrecht' in A Laufs, C Katzenmeier and V Lipp (eds), *Arztrecht* (8th ed. 2021) recital 12.
[125] U Wiesing, *Ethik in der Medizin* (2nd ed. 2004) 75.
[126] (Model) Professional Code for Physicians in Germany – MBO-Ä 1997 – The Resolutions of the 121st German Medical Assembly 2018 in Erfurt as amended by a Resolution of the Executive Board of the German Medical Association 14/12/2018 (hereafter MBO-Ä 1997).
[127] E.g. § 203 StGB (German Criminal Code) which protects patient confidentiality.
[128] Civil law regulates the contracts for the treatment of patients in §§ 630a ff. BGB.
[129] Cf. § 630a BGB, C Katzenmeier, 'BGB § 630a' in *BeckOK BGB* (61st ed. 2022) para. 1 *et seq.*
[130] M Quaas, '§ 14 Die Rechtsbeziehungen zwischen Arzt (Krankenhaus) und Patient' in R Zuck, T Clemens, and M Quass (eds), *Medizinrecht* (4th ed. 2018) recital 128.

influence the applicable medical standard on a national level. Overall, the WMA's guidelines would have a spillover effect in national professional law, whether in the area of professional regulations or in the scope of application of other federal or state laws. In this way, the contents of the guidance defined by the WMA could ultimately become binding for the individual physician licensed in Germany.

The situation is similar in Spain. The Spanish Medical Colleges Organization is a member of the WMA as the national medical association of Spain and 'regulates the Spanish medical profession, ensures proper standards and promotes an ethical practice'.[131] Furthermore, the WMA is the main instrument for the participation of national medical associations in international issues. For example, the American Medical Association, as a member of the WMA, makes proposals for international guidelines and agendas and lobbies at the national level to achieve the goals of physicians in the health field.[132]

IV. CONCLUSION: NECESSITY OF REGULATION BY THE WORLD MEDICAL ASSOCIATION

In order to close the gaps in the international guidance on the application of AI in medical care, active guidance by the WMA is recommended. Although it is not a subject of international law, meaning its guidance does not have legally binding effects, it is the only organization that has a strong indirect influence on national medical professional law through its members, as shown above. The incorporation of the contents of the guidance decided by the WMA is faster and less complex in this way than via the path of achieving legal effects through international soft law documents, particularly as the integration of the WMA guidelines into national professional laws reaches the physician actors that apply emerging technologies such as AI in only a few steps of implementation.

Furthermore, national professional laws and national professional regulations form not only the legal but also the ethical basis of the medical profession.[133] Consequently, professional law cannot be seen independently of professional ethics; instead, ethics constantly affect the legal doctor–patient relationship.[134] For example, the preamble to the German Model Code of Professional Conduct of the German Medical Association[135] states, among other things, that the purpose of the code of professional conduct is to preserve trust in the doctor–patient relationship, to ensure the quality of medical practice, to prevent conduct unbecoming a doctor, and to preserve the freedom of the medical profession. Furthermore, §2(1) sentence 1 MBO-Ä requires that physicians practice their profession according to their conscience, the prescriptions of medical ethics, and humanity. In addition, § 3(1) MBO-Ä also prohibits the practice of a secondary activity that is not compatible with the ethical principles of the medical profession. Preceding the regulations and the preamble of the model professional code of conduct is the medical vow set out in the WMA's Declaration of Geneva[136], which is a modernized form of the

[131] For more information see Organizatión Médica Colegial de España, 'Funciones del CGCOM' www.cgcom.es/funciones.

[132] For more information see American Medical Association, 'AMA's International Involvement' www.ama-assn.org/about/office-international-relations/ama-s-international-involvement.

[133] Bundesärztekammer, '(Muster-)Berufsordnung-Ärzte' https://www.bundesaerztekammer.de/themen/recht/berufsrecht.

[134] BVerfGE, 52, 131 (BVerfG *BvR* 878/74) para 116.

[135] MBO-Ä 1997 (n 126).

[136] WMA, 'Declaration of Geneva (1947), last amended by the 68th General Assembly in Chicago, USA, October 2017' (*WMA*, 9 July 2018) www.wma.net/policies-post/wma-declaration-of-geneva/.

Hippocratic Oath, itself over 2,000 years old. Altogether, this shows that ethics of professional conduct are not isolated from the law; they have a constant, universal effect on the legal relationship between the physician and the patient. Since the law largely assumes as a legal duty what professional ethics require from the physician,[137] the inclusion of medical ethics principles in professional law seems more direct in its effect than the inclusion of bioethical principles in international soft law.[138]

From this example and the overall impact of the Declaration of Helsinki, it is clear that the WMA has the potential to work toward a standard that is widely recognized internationally. The orientation of the WMA towards European or American medical ethics must, however, be kept in mind when issuing guidelines. In particular, the ethical concerns of other members should be heard and included in the internal deliberation. Furthermore, the associations of other medical professions, such as the International Council of Nurses,[139] with whom partnerships already exist in most cases,[140] should be consulted, not least because their own professional field is strongly influenced by the use of AI in the treatment of patients, but also to aid the dissemination of medical ethics and standards throughout the health sector. Expanding participation in deliberation increases the legitimacy of the WMA's guidelines and thus the spillover effect into the national professional law of physicians and other professions beyond. A comparison with other international organizations, such as UNESCO, also shows that the WMA, precisely because it is composed of physicians and because of its partnerships with other professional organizations, is particularly well suited from a professional point of view to grasp the problems of the use of AI in medical treatment and to develop and establish regulations for dealing with AI in the physician–patient relationship as well as in the entire health sector.

[137] 'Far more than in other social relations of human beings, the ethical and the legal merge in the medical profession.' E *Schmidt*, 'Der Arzt im Strafrecht' in A Ponsold (ed), *Lehrbuch der gerichtlichen Medizin* (2nd ed. 1957) 1, 2; BVerfGE, 52, 131 (BVerfG *BvR* 878/74).

[138] UNESCO states, for example, that 'Human rights law contains provisions that are analogous to the principles that flow from analysis of moral obligations implicit in doctor–patient relationships, which is the starting point, for example, of much of the Anglo-American bioethics literature, as well as the bioethics traditions in other communities.' UNESCO IBC, 'Report on Human Gene Therapy' SHS-94/CONF.011/8, Paris, 24.12.1994, IV.1.

[139] International Council of Nurses www.icn.ch.

[140] WMA, 'Partners, WMA Partnerships' www.wma.net/who-we-are/alliance-and-partner/partners/.

23

"Hey Siri, How Am I Doing?"

Legal Challenges for Artificial Intelligence Alter Egos in Healthcare

Christoph Krönke

I. INTRODUCTION

In response to the question 'Hey Siri, how am I doing?', Apple's intelligent language assistant today only gives ready-made answers ('You're OK. And I'm OK. And this is the best of all possible worlds.'). In the foreseeable future, however, it is quite conceivable that intelligent systems with comprehensive access to the health data of individual users could provide information and assessments of an individual's state of health, make recommendations for a better way of life and possible treatments, and communicate directly with other actors in the medical field (e.g. a treating physician). This opens up the prospect that, with a simple touch of (or even a conversation with) our smartphones, we could enjoy all the promises generally associated with the digitalization of healthcare: comprehensive individual health data would be available and manageable anywhere and anytime, and they could be used to generate high-quality medical diagnoses using Artificial Intelligence (AI), such as those that are already within reach for skin cancer diagnosis[1] or breast cancer detection.[2]

At the same time, the perspective on AI Alter Egos in the health sector raises numerous legal questions. The most essential of these increasingly pressing issues shall be identified and briefly discussed in this contribution – in a way that is understandable not only for die-hard lawyers.[3] First and foremost, responsible AI Alter Egos in healthcare would certainly require, on the one hand, a high level of data protection and IT security, for example, with regard to an individual's informed consent to the data processing and with respect to the (centralized or decentralized) storage of health data. On the other hand, such dynamic systems would pose particular challenges to medical devices law, for instance with regard to the necessary monitoring of a self-learning system with medical device functions. Furthermore, conflicts of interest between the areas of law involved are becoming apparent, particularly with regard to the rather restrictive, limiting approach of data protection law on one side of the spectrum, and the rules of product safety law aiming for efficiency, high quality, and high performance of applications on the other

[1] There are already analytical methods for the detection of skin cancer that can be implemented using a commercially available smartphone and that are significantly more powerful than the cognitive abilities of the average dermatologist, cf. A Esteva and others, 'Dermatologist-Level Classification of Skin Cancer with Deep Neural Network' (2017) 542 *Nature* 115, 117 *et seq*.

[2] See e.g. ED Pisano, 'AI Shows Promise for Breast Cancer Screening' (2020) 577 *Nature* 35, 35 *et seq*.

[3] Many of the legal considerations I am making in this chapter are essentially based on my thoughts on data protection and medical devices law developed in my habilitation thesis, published as C Krönke, *Öffentliches Digitalwirtschaftsrecht* (2020) 467 *et seq*. (data protection law) and 500 *et seq*. (medical devices law).

side. With my considerations I would like to show that, all in all, the development of AI Alter Egos in healthcare will require an evolving interpretation of the applicable legal frameworks while – at the same time – ensuring that these systems make responsible decisions. Ignoring either of these necessities would put both the individual patient's (data) sovereignty and the quality of the system outputs at stake.

I would like to proceed as follows: first of all, I would like to outline and describe the functionalities of AI Alter Egos in the healthcare sector,[4] namely the functions of an Alter Ego as a program for storing and managing individual health data,[5] as a software for generating individual medical diagnoses,[6] and finally as an interface for a collective analysis and evaluation of Big Health Data.[7] On this basis, I will identify the key elements of the applicable legal framework and discuss the three basic functions of an AI Alter Ego in light of the basic requirements following from this framework.[8] In doing so, I will focus primarily on the supranational requirements of European Union law so as not to become entangled in the thicket of national legislation.[9]

II. AI ALTER EGOS IN HEALTHCARE: CONCEPTS AND FUNCTIONS

In determining the concept and the description of the aforementioned functions of an AI Alter Ego in the healthcare sector, I am guided primarily by the considerations of *Eugen Münch*[10] who has been developing the idea of a digital Alter Ego for decades[11]. This is mainly due to the fact that his ideas seem very sound and general and do not reflect a concrete business model, but rather the main features that any AI Alter Ego in healthcare could have. Moreover, *Münch* had anticipated much of what many digital assistants and smart objects are designed for today. In the context of this contribution, it should remain open whether the carrier of an Alter Ego in the healthcare sector should be one or more decidedly state players or (public or private) economic enterprises, and whether the Alter Ego can operate on the basis of a specific legal framework or on the general basis of private contracts.[12] Certainly, the past has shown that the innovative and performance capabilities of private sector players are often superior to those of digital government initiatives. Even if Alter Ego projects should initially come from the private sector, however, one thing must be clear from the outset: the overriding (ethical) principle behind the idea of an Alter Ego in the health sector is not to enable utmost economic usability of health data, but rather to preserve the data sovereignty of the individual.

[4] See Section II.
[5] See Section II 1.
[6] See Section II 2.
[7] See Section II 3.
[8] See Section III.
[9] For this reason, specific national legislation, such as the provisions of the 2019 Digital Supply Act (*Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation)* (Digitale-Versorgung-Gesetz, DVG) will not be covered. For more information on this legislation cf. J Kühling and R Schildbach, 'Die Reform der Datentransparenzvorschriften im SGB V' (2020) 2 *NZS* 41, 41 *et seq.*
[10] Founder of the *Münch Foundation*. See www.stiftung-muench.org/.
[11] See e.g. the report on *Eugen Münch*'s idea: A Seith 'Sanierung via Laptopmedizin' *Der Spiegel* (12 January 2005) www.spiegel.de/wirtschaft/landklinik-sterben-sanierung-via-laptopmedizin-a-387338.html. *Münch* recently appointed an informal 'Digital Alter Ego' expert commission, of which I have been a member since early 2020.
[12] These are highly significant *organizational* issues that are undoubtedly crucial to the success of any Alter Ego project. However, they depend on the political will and the specific legal framework of individual countries and therefore cannot be discussed in detail in this chapter.

This being said, the general idea of an AI Alter Ego in healthcare involves two components and key functions: database functions and diagnostic functions.

1. Individual Health Data Storage and Management

The prerequisite for AI Alter Egos is a vast database that contains and manages as much personal health data of individual users as possible. In the ideal case, the entire individual data stock forms and reflects a digital image of the physical condition of the individual – in other words, a (complete) digital 'Alter Ego'. In this way, the individual user has (at least theoretically) full access to the health-related information relating to him or her and can grant third parties, such as physicians, health companies, or insurances, access to a specific or several data areas too; subject, of course, to the practically, highly critical question of suitable data formats and interfaces. From a purely technical point of view, storage of the health data of all Alter Egos in a central database is just as conceivable as decentralized storage on systems that are controlled by the individual users or trustworthy third parties. However, as has been stated at the outset, the Alter Ego is designed as a tool that is intended to serve, first and foremost, as a benefit to the user. It shall, therefore, enable him or her to decide independently and responsibly ('sovereignly') on the access to and use of his or her health data. This idea of the individual's health-specific 'data sovereignty' can hardly be reconciled with a central storage of his or her data – let alone with an outsourcing in 'health clouds' located beyond European sovereign borders.

2. Individual Medical Diagnostics

Building on this storage and management function, the digital Alter Ego should also have the potential to generate customized and high-quality medical diagnoses, taking into account all available health-related data points of the individual, possibly monitored on a real-time basis. When classifying this second, diagnostic function, however, one should follow a strict sense of reality. On the basis of the common differentiation, to be thought of on a sliding scale, between 'weak' (or 'narrow') AI, which is merely involved in the processing of concrete, relatively limited tasks, and 'strong' (or 'general') AI, which can be entrusted with comparatively comprehensive tasks like a human doctor,[13] all of the intelligent diagnostic systems that are, will, or might be implemented in the foreseeable future can be clearly classified as forms of narrow AI, with very specific functions such as cloud-based applications that analyze and interpret computed tomography (CT) images using self-learning algorithms to prepare medical reports[14]. Strong intelligent systems, on the other hand, are the stuff for science fiction novels and movies and should therefore not be the basis for legal considerations.

[13] Cf. for this differentiation for instance I Revolidis and A Dahi 'The Peculiar Case of the Mushroom Picking Robot: Extra-contractual Liability in Robotics' in M Corrales, M Fenwick, and N Forgó (eds), *Robotics, AI and the Future of Law* (2018), 57–59; see also the differentiation made in the AI strategy of the German Federal Government: Die Bundesregierung, 'Strategie Künstliche Intelligenz der Bundesregierung' (KI Strategie Deutschland, November 2018) 4, 5 https://www.bmbf.de/bmbf/shareddocs/downloads/files/nationale_ki-strategie.pdf?__blob=publicationFile&v=1.

[14] In 2019, for example, the Siemens AI-based AI-Rad Companion Chest CT program was the first application of the company's AI-Rad Companion platform to receive CE marking (see M Bludszuweit, 'KI-basierte Software AI-Rad Companion Chest CT von Siemens Healthineers für Europa zugelassen' (Siemens Healthineers, 26 July 2019) www.siemens-healthineers.com/de/press-room/press-releases/pr-20190726028shs.html). The program evaluates CT images of the thorax from any source, highlights abnormalities with respect to the corresponding organs (heart or lung), the carotid artery and vertebrae, and automatically generates a report for the radiologist, including any indications of possible abnormalities.

3. *Interface for Collective Analysis and Evaluation of Big Health Data*

The performance of the diagnostic functions depends on the quantity and quality of the health data, on the basis of which the algorithms used in the Alter Ego are trained and ultimately formed into robust decision rules. Against this background, a possible third, rather secondary function of the digital Alter Egos in their entirety could be to provide an all-encompassing data basis for its various possible diagnostic functions. In this respect, the individual Alter Ego could be both the limiting and enabling interface for a supra-individual (collective) analysis and evaluation of Big Health Data, from which the individual 'data sovereign' could ultimately benefit. Even if this function is reminiscent of the dystopian scenario in which humans merely act as data sources and mutate into 'transparent patients' – the price of any medical evaluation method, however advanced, is always the availability of a comprehensive basis of health data.

III. KEY ELEMENTS OF THE LEGAL FRAMEWORK AND LEGAL CHALLENGES

As explained in the introduction, the legal framework for the establishment and operation of digital Alter Egos is primarily provided by European data protection law[15] and the law on medical devices.[16] In the following, I will put each of the aforementioned functions of an Alter Ego against the background of these legal rules and assess the prospect of AI Alter Egos in healthcare under the existing legal framework. In doing so I will focus on the scope of application as well as the material goals and basic concepts of these regimes.

1. *European Data Protection Law*

In order to adequately assess the specific data protection standards in their relevance for Alter Egos, it is not sufficient to make general references to the protection of informational self-determination or the rights to privacy and the protection of personal data.[17] As a matter conceived in terms of 'risk law',[18] data protection law shields the rights and interests of the persons concerned from various risks that can be typified to a certain extent. The resulting need for protection forms the actual concrete purposes of data protection law. The processing of personal data by digital Alter Egos touches on several of these purposes, which, in turn, can be assigned to the two fundamental protection concepts of data protection law, namely, the limitation and transparency of data processing.[19] Taking account of the different basic functions of AI Alter Egos, the following major data protection goals can be distinguished in the context of AI Alter Egos in healthcare.

[15] See Section III 1.

[16] See Section III 2. The applicable Medical Devices Regulation will be supplemented in the foreseeable future by the EU Artificial Intelligence Act, which at least in its draft version (see COM(2021) 206 final) refers to the Medical Devices Regulation and modifies it slightly with regard to high-risk systems.

[17] See the Charter of Fundamental Rights of the European Union (26 October 2012) 2012/C 326/02 (Charter of Fundamental Rights), Articles 7 and 8.

[18] The characterization of data protection law as a risk-focused legal regime seems not to be controversial, even though it is rarely explicitly addressed – see as an exception for example K Ladeur, 'Das Recht auf informationelle Selbstbestimmung: Eine juristische Fehlkonstruktion?' (2009) 62 *DÖV* 45, 53 et seq.

[19] Cf. with reference to the distinction of (limiting) opacity tools and (transparency-creating) transparency tools by P De Hert S Gutwirth 'Regulating Profiling in a Democratic Constitutional State' in E Claes, S Gutwirth, and A Duff (eds), *Privacy and the Criminal Law* (2006) 67 *et seq.*; N Marsch, *Das europäische Datenschutzgrundrecht* (2018) 96 *et seq.*, who refers to these concepts as 'protection goals'.

a. Limitation of Data Processing: Data Protection-Friendly and Secure Design

The individual data storage and management functions of Alter Egos easily activate the data protection requirements under both the General Data Protection Regulation (GDPR)[20] and the supplementary European basic rights on data protection.[21] All health-related information relating to individuals is personal data – even particularly sensitive in the sense of Article 9 of the GDPR – and all possible 'work steps' of data handling by the Alter Ego are subject to the processing operations defined in Article 4(2) of the GDPR, such as the collection, storage, reading, querying, matching, use, modification, and transmission of personal data.

Additionally, with regard to the function of Alter Egos as interfaces to a collective database for a comprehensive analysis and evaluation of Big Health Data, the data protection rules are likely fully applicable as well. In the context of medical treatments, almost every piece of information can be assigned a personal and health reference that makes the person behind it at least 'identifiable' in the sense of Article 4(1) GDPR. In particular, medical data like a large blood count or an ECG recording are so unique to an individual that they can hardly be fully anonymized. Complete technical anonymization, which would lead to the inapplicability of data protection law, is therefore illusory. In this respect, it is certainly true that, in principle, 'anonymous data' no longer exists in the healthcare sector.[22]

The data protection rules of the GDPR will thus subject almost every single processing of health-related data in Alter Egos to certain requirements with regard to the 'whether' and 'how' of data processing. With regard to the 'whether' of lawful data processing, Article 6(1) GDPR establishes the principle that processing of personal data is only permissible if it can be based on one of the processing situations mentioned in Article 6(1)(a) to (f) GDPR (the so-called prohibition principle). In particular, Articles 6(1)(a) and 9(2)(a) as well as Articles 6(1)(e) and 9(2)(g) and (h) of the GDPR can be considered as the predominant legal basis for the processing of health data by an Alter Ego, since the processing operations would be regularly based either on the explicit consent of the users or on specific legal provisions introduced by Member States in order to create a legal basis for the storage, management, and diagnostic analysis of individual health data. In addition, the opening clause of Article 9(2)(j) GDPR may also become relevant specifically for collective analysis and evaluation. This allows Member States to create legal processing powers for 'scientific research purposes' to a large extent, including also private research.[23] This legitimizes researchers to process health data even without the consent of the data subjects. Despite all the emphasis on the high level of protection in the health sector, the GDPR thus gives research interests comprehensive priority over the data protection interests of the data subjects.

[20] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

[21] See in particular Articles 7 and 8 of the Charter of Fundamental Rights.

[22] Cf. M Martini and M Hohmann 'Der gläserne Patient: Dystopie oder Zukunftsrealität? Perspektiven datengetriebener Gesundheitsforschung unter der DS-GVO und dem Digitale-Versorgung-Gesetz' (2020) 49 *NJW* 3573, 3574 (hereafter Martini and Hohmann, 'Der gläserne Patient'). Due to this lack of watertight anonymization possibilities *de facto*, they plead for the introduction of a concept of *legal* anonymization *de lege ferenda*, which would eliminate the identifiability of a data subject through health data by legal fiction, as long as sufficient technical and organizational security measures were in place.

[23] It should be noted that this (wide) interpretation of the term 'research' is disputed in legal scholarship. Some authors would like to interpret Art. 9 GDPR as exclusively referring to research in the public interest, see e.g. T Weichert 'Art 9 Verarbeitung besonderer Kategorien personenbezogenere Daten' in J Kühling and B Buchner (eds), *Datenschutz-Grundverordnung Bundesdatenschutzgesetz: DS-GVO/BDSG* (2nd ed. 2018) para 122. For a view similar to the one taken in this contribution cf. for instance, Martini and Hohmann, 'Der gläserne Patient' (n 22) 3576.

With regard to the 'how' of lawful data processing, Article 5 GDPR defines the essential 'principles of data processing', which include in particular the principles of purpose limitation,[24] data minimization[25] and storage limitation[26]. In addition to these basic processing rules, the Union's data protection legislation contains numerous other provisions. Some of these supplement the basic rules with sector-specific requirements, for example, with the particularly strict requirements for the processing of health-related data pursuant to Article 9 GDPR. Others specify, concretize, and flank them in more detail, for example in the rights of data subjects pursuant to Article 12 et seq. GDPR, and in some cases they do so by adding structural requirements beyond concrete data processing, like by requiring data protection-friendly and secure technology design in accordance with Article 25(2) and Article 32 GDPR.

In more concrete terms, the principle of purpose specification and limitation under Article 5 (1)(b) GDPR requires that the information be collected only 'for specified, explicit and legitimate purposes' and 'not further processed in a way incompatible with those purposes'. The importance of this principle is underlined by its embodiment in Sentence 1 of Article 8(2) of the Charter of Fundamental Rights. Therefore, the storage of health and other personal data 'for undetermined and not yet determinable purposes' is clearly impermissible under European Union law.[27] Otherwise, the data subjects would no longer be able to see by which bodies the specifically collected personal data are processed in which context. The principle of purpose limitation is supplemented by the principles of data minimization and necessity under Article 5 (1)(c) GDPR. Accordingly, the collection and storage of each piece of information must be necessary in relation to the specified processing purposes, in other words, it must be necessary for the specified diagnostic and other medical purposes. In the case of health-related information of a particularly sensitive nature, the need for data collection may be condensed into a specific decision to be taken.

Against this background, any storage of health data would have to be carried out for a definable medical purpose from the outset. The monitoring of bodily functions 'into the blue', that is, for yet unknown medical purposes that might (or might not) become relevant in the future, seems inadmissible. The creation of a 'digital Alter Ego' in the sense of a complete image of all physical processes in the patient's body, irrespective of an existing medical need, is therefore hardly possible under current data protection law – at least at first glance.

The specific requirements that can be derived from the principle of purpose limitation and the principle of necessity and data minimization continue to apply when accessing and retrieving information stored in the Alter Ego. For example, the principle of purpose limitation prohibits the processing of stored data for purposes that are not compatible with the originally defined purpose of collection. Accordingly, changes of purpose with regard to the processing of health-related data are only permissible if the conditions set out in Article 6(4) GDPR are met. Thus, either the (explicit) consent of the data subject is obtained[28] or another reason pursuant to Article 9(2) GDPR is available, in which case an additional compatibility check is to be carried out in accordance with Article 6(4) GDPR additionally.[29]

[24] Article 5(1)(b).
[25] Article 5(1)(c).
[26] Article 5(1)(d).
[27] Cf. (in a different, public context) CJEU, Joined Cases C-293/12 and C-594/1 *Digital Rights Ireland Ltd v Minister for Communications and Others* (8 April 2014).
[28] See GDPR, Article 9(2)(a).
[29] For a detailed analysis of the requirements following from GDPR, Article 6(4) see e.g. B Buchner and T Petri 'Art 6 Raeumlicher Anwendungsbereich' in J Kühling and B Buchner (eds), *Datenschutz-Grundverordnung Bundesdatenschutzgesetz: DS-GVO/BDSG* (3rd ed. 2020) paras 178 *et seq*.

Such changes of purpose will likely become inevitable with the increasing use of Alter Egos as well as the extension of their diagnostic function. One could think of information initially collected and stored solely for the purpose of monitoring cardiovascular functions that is later being processed for the purpose of cancer detection, too. As long as the general medical purpose of data processing is not abandoned, the compatibility test for both individual diagnostic and collective analysis and evaluation purposes is in general complied with; provided an interpretation taking the individual's interest in the performance of his or her own Alter Ego into account is carried out. However, this performance depends crucially on the fact that health data which were initially collected in a permissible manner can also be processed for additional purposes, including the generation of decision rules on the basis of large supra-individual (big data) databases. With regard to general research purposes, this idea has been explicitly laid down in the GDPR: according to Article 5(1)(b) GDPR, processing for (further) scientific research purposes is 'not considered incompatible with the original purposes'. This flexibilization of the purpose limitation principle does not exempt the person responsible from checking the compatibility of the secondary purpose with the primary purpose according to Article 6(4) GDPR on a case-by-case basis, the principle of purpose limitation is still valid – as a rule, however, he may assume that compatibility is guaranteed.[30]

Most certainly, the conception of a comprehensive individual health database, which can also form the foundation for potential collective (Big Health) data analysis and evaluation processing, involves highest structural dangers and risks with respect to both the lawfulness of the processing and the security of the stored information.[31] Automated processing of health data and the accessing of these data (both on the basis of centralized and decentralized storage system) entail a particular risk of inadmissible or even abusive input and accessing. This is in obvious tension with the requirements in Articles 24 and 25(1) GDPR, according to which the responsible body must take 'appropriate technical and organisational measures', taking into account the relevant risks, which serve to 'implement data protection principles, such as data minimisation, in an effective manner and to integrate the necessary guarantees in the processing in order to meet the requirements of this Regulation and protect the rights of the data subjects'. Similar structural requirements are laid down in Article 32 GDPR specifically with regard to data security.[32]

[30] Cf. A Roßnagel, 'Datenschutz in der Forschung' (2019) 4 *ZD* 157, 162.

[31] It should be mentioned that the field of 'data protection and Big Data' has become a subject of extensive research and will, as such, not be further discussed here. See e.g. T Weichert 'Big Data und Datenschutz – Chancen und Risiken einer neuen Form der Datenanalyse' (2013) 6 *ZD* 251; A Roßnagel, 'Big Data – Small Privacy? Konzeptionelle Herausforderungen für das Datenschutzrecht'(2013) 11 *ZD* 562, 562 *et seq*.; JP Ohrtmann and S Schwiering, 'Big Data und Datenschutz – Rechtliche Herausforderungen und Lösungsansätze' (2014) 41 *NJW* 2984, 2984 *et seq*.; T Helbling, 'Big Data und der datenschutzrechtliche Grundsatz der Zweckbindung' (2015) 3 *K&R* 145, 145 et seq.; P Richter, 'Datenschutz zwecklos? – Das Prinzip der Zweckbindung im Ratsentwurf der DSGVO' (2015) 39 *DuD* 735, 735 *et seq*.; C Werkmeister and E Brandt, 'Datenschutzrechtliche Herausforderungen für Big Data' (2016) 4 *CR* 233, 237 *et seq*.; K Ladeur, '"Big Data" im Gesundheitsrecht – Ende der Datensparsamkeit?"' (2016) 40 *DuD* 360, 360–361; N Culik and C Döpke, 'Zweckbindungsgrundsatz gegen unkontrollierten Einsatz von Big Data Anwendungen – Analyse möglicher Auswirkungen der DS-GVO' (2017) 5 *ZD* 226, 228; T Hoeren, 'IT- und Internetrecht – kein Neuland für die NJW' (2017) 22 *NJW* 1587, 1591; BP Paal and M Hennemann, 'Wettbewerbs- und daten(schutz) rechtliche Herausforderungen' (2017) 24 *NJW* 1697, 1700 *et seq*.; see also the contributions of G Hornung, 'Erosion traditioneller Prinzipien des Datenschutzrechts durch Big Data' and Y Hermstrüwer, 'Die Regulierung der prädikativen Analytik: eine juristisch-verhaltenswissenschaftliche Skizze' in W Hoffmann-Riem (ed), *Big Data – Regulative Challenges* (2018) 79, 99.

[32] The relationship between GDPR, Article 32 and Article 24 *et seq*. DSGVO is illuminated by M Martini in BP Paal and DA Pauly (eds), *Datenschutz-Grundverordnung Bundesdatenschutzgesetz: DS-GVO/BDSG* (2nd ed. 2018) paras 7 *et seq*.

In view of the obligations to ensure that technology is designed in a 'privacy by design' manner, it is imperative for any healthcare Alter Ego system that a highly effective access rights management system be introduced that is absolutely subordinate to the 'health data sovereignty' of the individual. Furthermore, in view of the high risks involved, it is likely to be imperative to develop a decentralized (rather than a centralized) data storage system. Against this background, the ethical principle of data sovereignty of the individual also forms a legal principle with binding organizational effects for any Alter Ego in healthcare.

b. Securing a Self-Determined Lifestyle and Protection from Processing-Specific Errors through Transparency

In contrast to its database functions, the diagnostic function of an AI Alter Ego rather faces the typical data protection objectives that apply to all intelligent AI systems. Especially, the specific lack of transparency of algorithmically controlled decisions of intelligent systems challenges the goal of guaranteeing an autonomous self-determined lifestyle. An example with special relevance to data protection law is medical diagnoses that are made according to rules based on Big Data procedures. These decisions are typically based firstly on correlations (and thus not necessarily on causalities) and secondly on a multitude of different health-related data in the context of the concrete decisions. The results of the medical recommendations of an Alter Ego in the healthcare sector could range from the (comparatively harmless) recommendation to take a walk to stimulate the circulation to more sensitive predictions such as suspected sugar disease or a skin cancer diagnosis. If the rules and factors relevant to the decision in question, particularly with regard to the relevance of certain health-related and other personal circumstances, are not sufficiently clear to the person affected by the decision, this person has, on the one hand, no opportunity to adjust his or her behavior to the decision and, on the other hand, cannot recognize or correct factual errors of the Alter Ego.[33] In such a context, an autonomous, self-determined way of life appears to be possible only to a limited extent as the range of diagnostic possibilities increases. For such reasons, the creation of transparency in data processing has long been a recognized principle of data protection law.[34] The diagnostic function of an Alter Ego operating by means of AI is, therefore, in a specific tension between this principle and the many transparency-securing provisions of data protection law.

Furthermore, the use of intelligent systems such as AI Alter Egos in healthcare regularly touches on the need to protect the data subject from processing operations based on inappropriate decision rules. For example, if the decisions fail to achieve their medical (data processing) purpose due to inappropriate programming or use of the Alter Ego, they might generate inappropriate output. On the one hand, this addresses the possible specific quality problems of intelligent systems in general.[35] These problems can be based on various factors, such as the inferiority of the data basis used for the development of the decision rules, the improper or even illegal programming of the Alter Ego, or its use in a context that is not suitable for it. On the other hand, a specific element of the regulatory objective of avoiding inappropriate output of data processing lies in the protection against discrimination specific to data processing. What is meant is not unequal treatment as such, which occurs when a person is discriminated against based on particularly sensitive personality traits such as origin or disability. Rather, it refers to

[33] Cf. M Martini, *Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz* (2019) 30 *et seq*.

[34] See GDPR, Article 5(1).

[35] Cf. for example T Wischmeyer, 'Regulierung intelligenter Systeme' (2018) 143 *AöR* 1, 23 *et seq*. who also treats quality control as an overarching regulatory concern and protection against discrimination as a special problem of 'failure' of intelligent systems.

more disadvantageous treatment in a broader sense; this is when the person concerned belongs to a group of persons previously formed by the system. This second definition includes circumstances in which persons are assigned to a group that was defined specifically for one person by the system in the first place. Therefore, such groups can be understood as 'tailor-made'.

The decision-making rules of an Alter Ego in the health sector will typically be based on the linking of certain health or other personal data points, like name, place of residence, educational level or income, eating, and other habits. These data points are often 'developed' by the system itself and typically include the results expected from the output of the Alter Ego, such as a specific diagnosis of a disease or general life expectancy. Even though Big Data procedures in particular aim to achieve the most granular classifications and evaluations by including as many data points as possible, these procedures inevitably lead to the formation of groups of people and a certain expectation or evaluation. To provide an example: higher risk of suffering from a certain disease might be linked to the affiliation to a certain group profile, for instance, people with a foreign name, a place of residence with low purchasing power, an unhealthy diet, moderate exercise, no university studies, etc. Because the Alter Ego does not necessarily include all individual health-related characteristics of a person and rather decides merely on random group membership based on more or less health-related (and other personal) data, a negative decision for the person with the desired characteristic (like low risk of illness) contrary to the system expectation based on his or her profile may prove to be arbitrary.[36]

One aspect however must be particularly emphasized at this point, as it is often not sufficiently taken into account in legal scholarship:[37] data protection law itself does not prohibit incorrect or unlawful outputs, and in particular it does not prohibit general discrimination. The fact that unequal treatment based on gender, origin, other group memberships, or simply arbitrariness is not permissible does not follow from data protection regimes, but rather from substantive anti-discrimination legislation. Only the *structural bias* of automated data processing in general and of intelligent Alter Egos in particular is relevant from a perspective of data protection law. Such structural biases include the tendency to treat individuals in relation to a specific (medical) processing purpose on the basis of selective, typifying characteristics and this treatment being potentially inappropriate, arbitrary, and/or contrary to the purpose of the processing.

2. *European Medical Devices Regulation*

In the healthcare sector, such substantial-qualitative normative requirements – which cannot be derived from data protection law itself – arise from European medical devices law with regard to the outputs of an AI Alter Ego. According to the two introductory recitals of the applicable Medical Devices Regulation (MDR),[38] European medical devices law not only aims to ensure a functioning internal market for medical devices and thus pursues both cross-border coordination and economic promotion purposes, it is also supposed to guarantee high standards with regard to the quality (performance of the products) and safety (prevention of hazards and risks) of medical devices. First of all, it depends on the medical device legal classification of the individual

[36] Cf. with regard to AI-based decisions in general M Martini, *Blackbox Algorithmus* (n 33) 50.

[37] See for the following considerations C Krönke, *Öffentliches Digitalwirtschaftsrecht* (2020) 500 *et seq.*

[38] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, OJ 2017 L 117 (MDR).

functions of an AI Alter Ego[39] whether and to what extent the general objectives and the specific requirements of MDR[40] apply.

a. Classifying AI Alter Ego Functions in Terms of the Medical Devices Regulation

It goes without saying that software like an AI Alter Ego or, more precisely, individual functions of it can be classified as 'medical devices' in the legal sense. Software and software-supported products have been playing a significant role in the markets for medical services in the broader sense for some time. Possible distribution channels include software purchase or software rental as well as purely remote sales-based diagnostic or therapeutic services.[41] Possible applications which could also be used as part of an Alter Ego system range from comparatively simple computer programs, such as classical practice software for maintaining electronic patient records or health-related smart watch functions[42], to more complex, intelligent programs and systems, such as cloud-based applications that analyze and interpret computed tomography (CT) images using self-learning algorithms to prepare medical reports.[43] A differentiation between different types of applications is particularly useful with regard to the respective use context intended, as the distinction between medical devices and non-medical devices as well as the classification according to different risk classes[44] is primarily based on the intended purpose of the product.[45] Against this background, four types of software functions can be distinguished from the outset in the context of AI Alter Egos in healthcare: (1) functions that qualify as 'software as a medical device' (so-called stand-alone software or software as a medical device – SaMD) and as (2) software as an accessory of a medical device; furthermore, Alter Ego functions that fall within the category of a (3) software as a component of a medical device (so-called integrated software), and finally (4) functions that merely qualify as software in the medical field.[46]

First of all, (1) certain Alter Ego functions could fall under the term 'medical devices' in themselves, if they are intended to fulfil one of the 'specific medical purposes' mentioned in Article 2(1) MDR, i.e. if they are intended to diagnose, monitor or treat diseases, injuries or disabilities. A direct effect in or on the human body is not necessary for this purpose; a provision 'for human beings' is sufficient, even if it is only aimed at indirect physical effect.[47] In this sense

[39] See Section III 2(a).

[40] See Section III 2(b).

[41] Such sales forms are also explicitly covered by medical devices law, see MDR, Article 6.

[42] For functions of the Apple Watch (so far in versions 4 and 5) there are CE markings for an 'ECG App', which records a 1-channel electrocardiogram (ECG) and evaluates it with regard to atrial fibrillation (AFib), as well as a function 'Messages in case of irregular heart rhythm', which analyses the pulse rate with regard to irregularities indicating AFib (see the description on www.apple.com/de/healthcare/apple-watch/).

[43] See the references earlier at (n 14).

[44] See MDR, Annex VIII 3(1): 'The application of the classification rules depends on the intended purpose of the products'.

[45] See the legal definition in Article 2(1) MDR, according to which each medical device 'shall fulfil one or more of the specific medical purposes [described in detail in the regulation]'.

[46] Cf. on this common classification, which is also the basis for the scheme of the Commission's Guidelines on the qualification and classification of stand-alone software used in healthcare within the regulatory framework of medical devices, European Commission DG Internal Market, Industry, Entrepreneurship and SMEs, 'Medical Devices: Guidance document' (2016) *MEDDEV* 2.1/6 9 *et seq*. (hereafter European Commission, 'Medical Devices'), for example R Oen, 'Software als Medizinprodukt' (2009) 2 *MPR* 55, 55 *et seq*.; M Klümper and E Vollebregt, 'Die geänderten Anforderungen für die CE-Kennzeichnung und Konformitätsbewertung auf Grund der Richtlinie 2007/47/EG' (2009) 2 *MPJ* 99, 100-101; S Jabri, 'Artificial Intelligence and Healthcare: Products and Procedures' in T Rademacher and T Wischmeyer (eds), *Regulating Artificial Intelligence* (2020) 307, 314 *et seq*.

[47] CJEU, C-329/16 *Snitem and Philips France* (26 January 2018) paras 27 *et seq* (herafter *Snitem and Philips France*).

(and explicitly according to the former directive terminology) 'independent'[48] software products are considered 'active' medical devices under Article (4) MDR, for which specific classification rules and material requirements apply; they are also subject to special regulations, such as those of the MDR's UDI[49] system). Practical examples of such SaMDs are decision-support programs comparing medical databases with the data of individual patients in order to provide medical personnel or patients directly with recommendations for the diagnosis, monitoring, or treatment of the patient in question.[50] The complex systems for the (possibly adaptive) analysis of image and other data with descriptive, predictive, or prescriptive functions mentioned earlier in this contribution also fall into this group of software products. This category is probably the most relevant for the diagnostic functions of an AI Alter Ego in healthcare.

Other Alter Ego functions will qualify as (2) 'accessories' in the sense of Article 2(2) MDR.[51] In contrast to (completely independent) standalone software, accessory software does not fulfil a specific medical purpose itself. However, it does fulfil such a purpose in combination with one or more other 'medical devices', by enabling or at least supporting its specific function as a medical device. In particular, software marketed separately for programming and controlling medical devices as well as their integrated software (e.g. of pacemakers)[52] is regularly qualified as accessory software. Against this background, support software that is compatible with an AI Alter Ego but marketed separately could fall within the category of an accessory.

Distinct from these first two categories are (3) supportive Alter Ego functions forming an integral part of one or more other Alter Ego functions that qualify as medical devices at the time of the placing on the market.[53] Important examples of such integrated software include programs for the control of medical devices, like blood pressure monitors[54] or the power supply.[55] Such programs are not treated as medical devices themselves but as mere components of the respective product.

In contrast, (4) all other functions of an AI Alter Ego would have – as such! – no relevance under medical devices law. These can be programs with essential but merely auxiliary functions such as collecting, archiving, compressing, searching, or transmitting data. Examples include important information and communication systems that are connected with the diagnostic functions of the Alter Ego such as communication systems for separate tele-medicine services,[56]

[48] Cf. critically with regard to the renouncement of this terminology in the MDR and the practical consequences of this renouncement UM Gassner, 'Software als Medizinprodukt – zwischen Regulierung und Selbstregulierung' (2016) 4 *MPR* 109, 110–111. The previous differentiation between independent and integrated software therefore should remain valid.

[49] Short for Unqiue Device Identification.

[50] See German Federal Office for Drugs and Medical Devices, 'Orientierungshilfe Medical Apps' (*BfArM*, 1 November 2015) https://docplayer.org/63901775-Bfarm-orientierungshilfe-medical-apps.html point 3 (hereafter BfArM, 'Orientierungshilfe Medical Apps'). Such a program was also the subject of the proceedings in CJEU, *Snitem and Philips France* (n 47) paras 17 *et seq*. After entering individual patient data, the program alerted the user to possible contraindications, interactions with other drugs and overdoses, etc.

[51] From recital 19 sentence 2 of the MDR it becomes clear that software can actually be accessories. This was previously controversial, see UM Gassner, 'Software als Medizinprodukt – zwischen Regulierung und Selbstregulierung' (2016) 4 *MPR* 109, 111.

[52] Cf. for this example M Klümper and E Vollebregt, 'Die geänderten Anforderungen für die CE-Kennzeichnung und Konformitätsbewertung auf Grund der Richtlinie 2007/47/EG' (2009) 2 *MPJ* 99, 100.

[53] Cf. for a general definition of 'integrated' medical software e.g. R Tomasini, *Standalone-Software als Medizinprodukt* (2015) 44.

[54] Cf. for this example G Sachs, 'Software in Systemen und Behandlungseinheiten' in UM Gassner (ed), *Software als Medizinprodukt – IT vs. Medizintechnik?* (2013) 31 *et seq*.

[55] M Klümper and E Vollebregt, 'Die geänderten Anforderungen für die CE-Kennzeichnung und Konformitätsbewertung auf Grund der Richtlinie 2007/47/EG' (2009) 2 *MPJ* 99, 100.

[56] Cf. BfArM, 'Orientierungshilfe Medical Apps' (n 46) point 3.

medical knowledge databases,[57] hospital information systems (HIS) with pure data collection, administration, scheduling, and accounting functions as well as picture archiving and communication systems (PACS) without reporting function[58]. Furthermore, as recital 19 sentence 1 of the MDR states in principle, programs used for lifestyle and well-being purposes are not sufficiently related to specific medical purposes. These include, in particular, the functions of a Smartwatch for recording and evaluating movement calories or sleep rhythm when using a lifestyle app. Of course, software with completely unspecific functions, for example operating systems or word processing program, are also irrelevant under medical devices law. Against the background of these considerations, software serving the individual data storage and management function of an AI Alter Ego as well as possible functions aiming for the collective analysis and evaluation of the (big) health data gathered through the participating Alter Egos in their entirety would – as such! – not qualify as 'medical devices' or 'accessories' under the MDR.

This does not mean, however, that the individual database functions and the collective Big Health Data functions of an AI Alter Ego are entirely irrelevant under medical devices law. It is not only the diagnostic functions being relevant. Of course, the usual case in practice[59] deals with information technology systems consisting of several modules. In such instances, some of these modules can be qualified typically as a medical device or accessory, while other modules can only be qualified as software in the medical field. Consequently, the rules of medical devices law, especially the obligation to label, only apply to the first-mentioned modules.[60] Nevertheless, it has probably become clear that the performance of the diagnostic functions of an AI Alter Ego is crucially dependent on the quantity and quality of the data sets, including the software used to store and manage, analyze, and evaluate them. Even if the databases and their management software as well as the algorithms used to analyze and evaluate them are not subject to medical devices law as such, their quality and design has a decisive influence on how the diagnostic functions are to be assessed under medical devices law. In this respect, the individual database functions and the Big Health Data functions of an AI Alter Ego are not directly, but indirectly relevant for the following medical devices law considerations.

b. Objectives and Requirements Stipulated in the MDR

The potentially high quantitative and qualitative performance of the diagnostic functions of AI Alter Egos affects the core objective of medical devices law to ensure high quality standards in the healthcare sector, just like the use of AI in the healthcare sector in general. The need for such systems including cost aspects becomes obvious if, for example, in a side-by-side comparison between 157 dermatologists and an algorithm for evaluating skin anomalies, only seven experts are able to make more precise assessments of skin abnormalities than the computer system.[61]

At the same time, the safety-related requirements of medical devices law are also touched upon. These requirements aim for the prevention and elimination of quality defects as well as imminent hazards and risks. The characteristic lack of transparency of algorithmic decision rules (which can produce unforeseen and unpredictable results) as well as the adaptability of continuously learning systems add specific risks to the increased basic risk inherent in all

[57] See CJEU, *Snitem and Philips France* (n 47) para 33.
[58] Cf. for the latter two examples again BfArM, 'Orientierungshilfe Medical Apps' (n 49) point 3.
[59] Cf. also with numerous practical examples in European Commission, 'Medical Devices' (n 46) 17, 18.
[60] See in principle CJEU, *Snitem and Philips France* (n 47) para 36.
[61] Cf. with this very example Y Frost, 'Künstliche Intelligenz in Medizinprodukten und damit verbunden medizinprodukte- und datenschutzrechtliche Herausforderungen' (2019) 4 *MPR* 117, 117.

medical devices. Yet, precisely this adaptability is considered particularly attractive in the field of intelligent medical devices. Nevertheless, and in view of the high-ranking fundamental rights to which medical device risks generally refer (life and limb), these specific risks must be taken seriously and addressed appropriately by the regulatory authorities.

Particularly relevant for the development and operation of Alter Egos in the health sector and their basic functions (i.e. indirectly for the individual database function and the collective Big Health Data function, directly for its diagnostic functions) are the structural requirements laid down by the MDR. A look at these structural requirements of medical devices law shows that the introduction of intelligent Alter Egos in the healthcare sector will encounter a legal matter that is already particularly well adapted to the specific technology-related risks of such products for the protected goods concerned.

At the top of structural requirements is the general obligation to ensure the safety and efficacy of the medical device,[62] which is differentiated by further requirements, such as the obligation to perform a clinical evaluation or a clinical trial according to Article 10(3) MDR.[63] For the marketing of intelligent Alter Egos, some of these specifications seem particularly relevant. For example, in addition to the obligation to set up a general quality management system as part of quality assurance, which has been customary for industrially producing companies for decades,[64] the MDR orders the introduction of a risk management system,[65] in the context of which the specific risks of software and data-based products in particular must also be explicitly addressed.[66] In addition, according to Article 10(10) MDR, the 'manufacturer' of the Alter Ego must set up a post-marketing surveillance system in the sense of Article 83 MDR. At least in theory, the typical possibility of unforeseen outputs of AI Alter Egos in general and the adaptability of continuous learning systems in particular can be countered with such systems. In accordance with the regulatory concept of medical devices law, these abstract and general requirements are also specified in more detail for software products by means of special ('harmonized') technical standards. Particularly relevant in this respect is the international standard IEC 62304[67], adopted by the responsible European standardization organization Cenelec, which supplements the risk management standard ISO 14971 with software-specific aspects and also formulates requirements for the development, maintenance, and decommissioning of stand-alone software and for integrated software.[68] In particular, these standards contain, for instance, guidelines for the handling of raw data and its transformation into 'clean data' as well as for the proper training and validation of algorithms.

It is quite likely that that new types of risks are created in the development of intelligent medical devices if AI Alter Egos became actually widely used and were replacing conventional medical services and institutions. Depending on whether and to what extent such scenarios

[62] MDR, Article 10(1) in conjunction with Annex I Chapter I 1.

[63] In addition to these general warranty and risk management requirements, there are also labeling, documentation, recording, reporting, and notification obligations that relate to the warranty and risk management requirements. For reasons of simplification, they will not be discussed further here.

[64] See MDR, Article 10(9) in connection with Annex IX Chapter I. Cf. on the emergence of quality assurance systems from the 1960s onwards and on the principles of quality management in detail F Reimer, *Qualitätssicherung. Grundlagen eines Dienstleistungsverwaltungsrechts* (2010) 115 *et seq.*

[65] MDR, Article 10(2) in conjunction with Annex I Chapter I 3.

[66] See MDR, Annex I Chapter II 17, in particular point 17.2 MDR: 'For products incorporating software or in the form of software, the software shall be designed and manufactured in accordance with the state of the art, taking into account the principles of software life cycle, risk management including information security, verification and validation'.

[67] International Standard IEC 62304 Medical Device Software – Software Life Cycle Processes.

[68] For further relevant standards, see for example the overviews in C Johner, M Hölzer-Klüpfel, and S Wittorf, *Basiswissen Medizinische Software* (2nd ed. 2015) 28 *et seq.*; G Heidenreich and G Neumann, *Software for medical devices* (2015) 260 *et seq.*

actually happen and, given the event that these new types of risks are not specifically addressed in the MDR or in other relevant harmonized standards, the corresponding standards can certainly be further developed. Manufacturers and 'notified bodies' (i.e. the certified inspectors of medical devices) are called upon to take account of the special features of intelligent systems in the context of conformity assessment by means of a risk-conscious but innovative interpretation of the regulatory requirements. Such an interpretative approach shall also be undertaken when such requires a specification or perhaps even a deviation of relevant technical standards.[69] It will be possible for instance, to derive certain Good Machine Learning Practices (GMLPs) from the general provisions of the MDR, including the reference to the development and production of software according to the 'state of the art'.[70] According to the GMLPs, for example, only training data suitable for the product purpose may be selected; training, validation, and test data must be carefully separated from each other, and finally, it is necessary to work towards sufficient transparency of the intended output and the operative decision rules.[71] Continuous Learning Systems in Alter Egos are systems with decision rules that can be continuously changed during product operation and therefore actually have AI in the narrower sense and their application may generate specific risks as well. In principle, a change in the decision rules can become legally relevant from three points of view: it can affect the performance, safety, or intended use and/or data input of the product or its evaluation.[72] The manufacturer has to prepare for such changes already under the current regulatory situation, especially since Article 83(1) and (2) MDR obliges him to monitor the system behavior in a way that is adequate for the risk and the product. The manufacturer will have to identify and address (by developing a specific algorithm change protocol) such expected changes already within the scope of the establishment of his risk management system (as pre-specifications).[73] In any case, the distribution of intelligent medical devices does not pose insurmountable difficulties for medical devices law.

However, against the backdrop of the 'general obligation to ensure the safety and efficacy of the medical device' as described and explained above, the restrictions imposed by data protection law on the collection, storage, management, and other processing of health-related information appear to be a possible point of conflict. If restrictions on the use of health-related data, such as limitations on the changes of purpose, prove to be an obstacle to the quality of

[69] A deviation then requires justification, see for example the explicit requirement in MDR, Annex IX Chapter I 2.3, which specifies the test program of an audit procedure by a Notified Body. Cf. on the delicate balance of technical standards between their function of concretizing legal norms on the one hand and the compulsion to design products in conformity with the standard on the other hand, which is to be avoided because it may not be appropriate to the risks and/or innovation, H Pünder, 'Zertifizierung und Akkreditierung – private Qualitätskontrolle unter staatlicher Gewährleistungsverantwortung' (2006) 5 *ZHR* 170 567, 571.

[70] See the formulation in MDR, Annex I Chapter I 17.2. If the harmonized standards do not (any longer) adequately reflect these requirements and a corresponding software product is assessed as compliant, the market surveillance authorities can nevertheless argue that the software product does not comply with the Regulation, as compliance with the standards pursuant to Art. 8 para. 1 MDR only gives rise to a presumption of conformity.

[71] For these examples of GMLPs, see the considerations at M Diamond and others, 'Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device' (FDA, 2019) www.fda.gov/media/122535/download 9–10 (hereafter Diamond and others, 'Proposed Regulatory Framework').

[72] These possible areas of change are already covered in the Medical Devices Regulation, namely in MDR, Annex VI Part C 6.5.2. Almost identical is the information given in M Diamond and others, 'Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device' (FDA, 2019) 6-7 www.fda.gov/media/122535/download, which differentiates between changes regarding performance, inputs and intended use.

[73] For such *SaMD Pre-Specifications* (SPS) and an *Algorithm Change Protocol* (ACP) see Diamond and others, 'Proposed Regulatory Framework (n 70) 10 *et seq.*

outputs for medical purposes, the question arises as to which regime should be given preference in case of doubt. Generalized statements are not helpful here. Rather, these problems should be handled on a case-by-case basis. Of primary relevance is the Alter Ego's concrete medical function specifically affected. In the context of particularly sensitive functions, quality problems or system failures can have particularly far-reaching or even fatal consequences; as in the monitoring of cardiovascular functions or in the diagnosis of serious diseases, any restrictions imposed by data protection law should be overcome by an appropriate interpretation of the legal bases of data protection law. Conversely, a function designed to encourage the data subject to take regular walks should not necessarily be able to access all information, especially highly sensitive information.

IV. CONCLUSION

Overall, my considerations have shown that Alter Egos in the health sector, while appearing somewhat futuristic, already have an appropriate legal framework – at least if it is handled in an appropriate manner that is open to development. The truism will apply: not everything that is technically possible will (immediately) be legally permitted. The creation of a completely 'transparent patient' is (rightly) forbidden in view of the data protection principles of purpose limitation, necessity, and data minimization. Instead, the creation of comprehensive individual health databases in Alter Egos must be carried out step by step. The argument that every health-related data could (in the future) have some kind of medical relevance does not hold water here. On the other hand, data protection law and its legal basis must be interpreted in a way that is open to development and innovation in order to enable medical services that are already feasible and to allow individuals to make comprehensive and effective use of their health data for medical purposes. In order to ensure the quality of these medical functions, the existing rules of medical devices law already provide appropriate instruments that can be easily and adequately applied to AI Alter Egos. Hence, if the existing legal requirements are handled correctly, a responsible and at the same time powerful use of AI Alter Egos in the health sector can go hand in hand.

24

'Neurorights'

A Human Rights–Based Approach for Governing Neurotechnologies

Philipp Kellmeyer

I. INTRODUCTION

The combination of digital technologies for data collection and processing with advances in neurotechnology promises a new generation of highly adaptable, AI-based brain–computer interfaces for clinical but also consumer-oriented purposes. By integrating various types of personal data – physiological data, behavioural data, biographical data, and other types – such systems could become adept at inferring mental states and predicting behaviour, for example, for intended movements or consumer choices. This development has spawned a discussion – often framed around the idea of 'neurorights' – around how to protect mental privacy and mental integrity in the interaction with AI-based systems. Here, I review the current state of this debate from the perspective of philosophy, ethics, neuroscience, and psychology and propose some conceptual refinements on how to understand mental privacy and mental integrity in human-AI interactions.

The dynamic convergence of neuroscience, neurotechnology, and AI that we see today was initiated by progress in the scientific understanding of brain processes, the invention of computing machines and algorithmic programming in the early and mid-twentieth century.

In his book *The Sciences of the Artificial*, computer science, cybernetics, and AI pioneer *Herbert A. Simon* characterizes the relationship between the human mind and the human brain as follows:

> As our knowledge increases, the relation between physiological and information-processing explanations will become just like the relation between quantum-mechanical and physiological explanations in biology (or the relation between solid-state physics and programming explanations in computer science). They constitute two linked levels of explanation with (in the case before us) the limiting properties of the inner system showing up at the interface between them.[1]

This description captures the general spirit and prevailing analogy of the beginnings and early decades of the computer age: just as the computer is the hardware on which software is implemented, the brain is the hardware on which the mind runs. In the early 1940s, well before the first digital computers were built, *Warren S. McCulloch* and *Walter Pitts* introduced the idea of artificial neural networks that could compute logical functions.[2] Later, in 1950, *Donald Hebb*

[1] HA Simon, *The Sciences of the Artificial* (2001) 83.

[2] WS McCulloch and W Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity' (1943) 5(4) *The Bulletin of Mathematical Biophysics* 115–133 https://doi.org/10.1007/BF02478259.

in *The Organization of Behavior*[3] developed a theory of efficient encoding of statistics in neural networks which became a foundational text for early AI researchers and engineers. Later yet, in 1958, *Frank Rosenblatt* introduced the concept of a perceptron, a simple artificial neural network, which had comparatively limited information-processing capabilities back then but constitutes the conceptual basis from which the powerful artificial neural networks for deep learning are built today.

Much of this early cross-fertilization between discoveries in neurophysiology and the design of computational systems was driven by the insight that both computers and human brains can be broadly characterized as information-processing systems. This analogy certainly has intuitive appeal and motivates research programs to this day. The aim is to find a common framework that unifies approaches from diverse fields – computer science, AI, cybernetics, cognitive science, neuroscience – into a coherent account of information processing in (neuro)biological and artificial systems. But philosophy, especially philosophy of mind, (still) has unfinished business and keeps throwing conceptual wrenches – in the form of thought experiments, the most famous of which is arguably *John Searle's* Chinese Room Argument[4] – into this supposedly well-oiled machine of informational dualism.

Today, through the 'super-convergence'[5] of digital and information technologies, this original affinity and mutual inspiration between computer science (artificial neural networks, cognitive systems, and other approaches) and the sciences of the human brain and cognition is driving a new generation of AI-inspired neurotechnology and neuroscience-inspired AI.[6]

In the field of brain–computer interfacing, for example, the application of AI-related machine learning methods, particularly artificial neural networks for deep learning, have demonstrated superior performance to conventional algorithms.[7] The same machine learning approach also excels in distinguishing normal from disease-related patterns of brain activity, for example, in finding patterns of epileptic brain activity in conventional electroencephalography (EEG) diagnostics.[8] These and other successes in applying AI-related methods to analysing and interpreting brain data drives an innovation ecosystem in which not only academic researchers and private companies, but also military research organizations invest heavily (and compete) in the field of 'intelligent' neurotechnologies.[9] This development has spawned an increasing number of analyses and debates on the ethical, legal, social, and policy-related relevance of

[3] DO Hebb, *The Organization of Behavior* (1949).

[4] JR Searle, 'Minds, Brains, and Programs' (1980) 3 *Behavioral and Brain Sciences* 417–457 https://doi.org/10.1017/S0140525X00005756.

[5] The confluence of big data, artificial neural networks for deep learning, the web, microsensorics, and other transformative technologies, *cf.* H Hahn und A Schreiber, 'E-Health' in R Neugebauer (ed), *Digital Transformation* (2019) 311–334 https://doi.org/10.1007/978-3-662-58134-6_19.

[6] P Kellmeyer, 'Artificial Intelligence in Basic and Clinical Neuroscience: Opportunities and Ethical Challenges' (2019) 25(4) *Neuroforum* 241–250 https://doi.org/10.1515/nf-2019-0018; AH Marblestone, G Wayne, and KP Kording, 'Toward an Integration of Deep Learning and Neuroscience' (*Frontiers in Computational Neuroscience*, 14 September 2016) 94 https://doi.org/10.3389/fncom.2016.00094.

[7] D Kuhner and others, 'A Service Assistant Combining Autonomous Robotics, Flexible Goal Formulation, and Deep-Learning-Based Brain–Computer Interfacing' (2019) 116 *Robotics and Autonomous Systems* 98–113 https://doi.org/10.1016/j.robot.2019.02.015; F Burget and others, 'Acting Thoughts: Towards a Mobile Robotic Service Assistant for Users with Limited Communication Skills' (*IEEE*, 9 November 2017) 1–6 https://doi.org/10.1109/ECMR.2017.8098658.

[8] LAW Gemein and others, 'Machine-Learning-Based Diagnostics of EEG Pathology' (2020) 220 *NeuroImage* 117021 https://doi.org/10.1016/j.neuroimage.2020.117021.

[9] P Kellmeyer, 'Big Brain Data: On the Responsible Use of Brain Data from Clinical and Consumer-Directed Neurotechnological Devices' (2018) 14 *Neuroethics* 83–98 https://doi.org/10.1007/s12152-018-9371-x (hereafter Kellmeyer, 'Big Brain Data'); M Ienca, P Haselager, and EJ Emanuel, 'Brain Leaks and Consumer Neurotechnology' (2018) 36 *Nature Biotechnology* 805–810 https://doi.org/10.1038/nbt.4240.

brain data analytics and intelligent neurotechnologies.[10] Central concepts in this debate are the notions of mental privacy and mental integrity.

In this chapter, I will first give an account of the current understanding as well as ethical and legal implications of mental privacy and propose some conceptual refinements. Then I will attempt to clarify the conceptual foundations of mental integrity and propose a description that can be applied across various contexts. I will then address the debate on neurorights and advocate for an intermediate position between human rights conservatism (no new rights are necessary to protect mental privacy and integrity) and human rights reformism (existing human rights frameworks are insufficient to protect mental privacy and integrity and need to be revised). I will argue that the major problem is not the lack of well-conceptualized fundamental rights but insufficient pathways and mechanisms for applying these rights to effectively protect mental privacy and mental integrity from undue interference.

II. MENTAL PRIVACY

1. *The Mental Realm: The Spectre of Dualism, Freedom of Thought and Related Issues*

As outlined in the introduction and in the absence of a universal definition, I propose the following pragmatic operational description: 'Mental privacy denotes the domain of a person's active brain processes and experiences – perceptions, thoughts, emotions, volition; roughly corresponding to *Kant's* notion of the *locus internus* in philosophy[11] – which are exceptionally hard (if not impossible) to access externally.' The mental 'realm' implicated in this description refers to an agent's phenomenological subjective experiences, indicated in language by terms such as 'thoughts', 'inner speech', 'intentions', 'beliefs', and 'desires', but also 'fear', 'anxiety' and emotions (such as 'sadness'). While it makes intuitive sense, from a folk-psychological perspective, calling for special protection to this mental realm is predicated on a precise understanding of the relationship between levels of subjective experiences and corresponding brain processes – a requirement that neuroscientific evidence and models cannot meet[12].

From a monist and materialist position, these qualitative terms offer convenient ways for us to refer to subjective experiences, insisting that there is – in the strict ontological sense – nothing but physical processes in the human body (and the brain most of all), no dualistic 'second

[10] P Kellmeyer and others, 'Neuroethics at 15: The Current and Future Environment for Neuroethics' (2019) 10(3) *AJOB Neuroscience* 104–110; S Rainey and others, 'Data as a Cross-Cutting Dimension of Ethical Importance in Direct-to-Consumer Neurotechnologies' (2019) 10(4) *AJOB Neuroscience* 180–182 https://doi.org/10.1080/21507740.2019.1665134; Kellmeyer, 'Big Brain Data' (n 9); R Yuste and others, 'Four Ethical Priorities for Neurotechnologies and AI' (2017) 551 (7679) *Nature News* 159 https://doi.org/10.1038/551159a (hereafter Yuste and others, 'Four Ethical Priorities for Neurotechnologies and AI'); M Ienca and R Andorno, 'Towards New Human Rights in the Age of Neuroscience and Neurotechnology' (2017) 13 *Life Sciences, Society and Policy* 5 https://doi.org/10.1186/s40504-017-0050-1 (hereafter Ienca and Andorno, 'Towards New Human Rights in the Age of Neuroscience and Neurotechnology').

[11] I Leclerc, 'The Meaning of "Space"' in LW Beck (ed), *Kant's Theory of Knowledge: Selected Papers from the Third International Kant Congress* (1974) 87–94 https://doi.org/10.1007/978-94-010-2294-1_10. This division into a *locus internus* (as described here) and *locus externus* – the set of externally observable facts about human behavior – is reflected in the ongoing debate about the nature of human phenomenological experience, consciousness, and free will in philosophy; the intricacies and ramifications of which lie outside of the scope of this article. For recent contributions to these overlapping debates, see e.g. the excellent overview in P Goff's *Galileo's Error* (2020).

[12] I deliberately refrain from qualifying this statement as to whether, and if so when, we should expect neuroscience to ever be able to give a full account of a mechanistic understanding, both for conceptual reasons and practical reasons, for example, inherent limitations of current, and likely future, measurement tools in observing brain processes at the 'right' levels of granularity or scale (microscale, mesoscale, and macroscale) and at the appropriate level of temporal and frequency-related sampling to relate them to any given subjective experience.

substance' or, as *René Descartes* referred to it, *mens rea*. In such an interpretation, there is no 'mind-body problem' because there is no such thing as a mind to begin with and the human practice of talking as if there was a mental realm that is separate from the physical realm arises from our (again folk-psychological, or anthropological) propensity to interpret our subjective experience as separate from brain processes, perhaps because we have no direct sensory access to these processes in the first place.

This spectre of dualism, the illusion – as a materialist (e.g. a physicalist) would put it – that our physical brain processes and our experiences are separate 'things', is so convincing and persuasive that it not only haunts everyday language, but is also deeply engrained in concept-formation and theorizing in psychological and neuroscientific disciplines such as experimental psychology or cognitive neuroscience as well as the medical fields of neurology, psychosomatic medicine, and psychiatry.[13]

To date, there is no widely accepted and satisfying explanation of the precise relationship between the phenomenological level of subjective experience and brain processes. This conundrum allows for a wide range of theoretical positions, from strictly neuroessentialist and neurodeterministic interpretations (i.e. there is nothing separate from brain processes; and brain activity does not give rise to but simply *is* nothing but neurophysiology), to positions that emphasize the '4E'[14] character of human cognition and all the way to modern versions of dualist positions, such as 'naturalistic dualism'[15]. An interesting intermediate position that has experienced somewhat of a renaissance in the philosophy of mind in recent years is the concept of panpsychism. The main idea in panpsychism is that consciousness is a fundamental and ubiquitous feature of the natural world. In this view, the richness of our mental experience could be explained as an emerging property that depends on the complexity of biological organisms and their central nervous systems.[16] Intriguingly, there seem to be conceptually rich connections between advanced neuroscientific theories of consciousness, particularly the so-called Integrated Information Theory (IIT)[17], and emergentist panpsychist interpretations of consciousness and mental phenomena.[18] The reason why this is relevant for our topic here – brain data, information about brain processes, and neurotechnology – is that these conceptual and neuroscientific advances in building a unified theory of causal mechanisms of subjective experience might become an important tenet for future analytical approaches to decoding brain data from neurotechnologies and inferring mental information from these analyses.

[13] Consider, for example, the concept of 'dissociation' in psychiatry (in the context of post-traumatic stress disorder) or neurology (in epilepsy), the notion that brain processes and mental processes can become uncoupled.

[14] The 4E framework emphasizes that human cognition cannot be separated from the way in which cognitive processes are embodied (in a physical body [German: '*Leib*']), embedded (into the environment), extended (how we use tools to facilitate cognition), and enactive (cognition enacts itself in interaction with others) R Menary, 'Introduction to the Special Issue on 4E Cognition' (2010) 9(4) *Phenomenology and the Cognitive Sciences* 459–463 https://doi.org/10.1007/s11097–010-9187-6.

[15] D Chalmers, 'Naturalistic Dualism' in S Schneider and M Velmans (eds), *The Blackwell Companion to Consciousness* (2017) 363–373 https://doi.org/10.1002/9781119132363.ch26.

[16] P Goff, *Consciousness and Fundamental Reality* (2017); P Goff, W Seager, and S Allen-Hermanson, 'Panpsychism' in EN Zalta (ed), *The Stanford Encyclopedia of Philosophy* (2020) https://plato.stanford.edu/archives/sum2020/entries/panpsychism/.

[17] G Tononi and others, 'Integrated Information Theory: From Consciousness to Its Physical Substrate' (2016) 17(7) *Nature Reviews Neuroscience* 450–461 https://doi.org/10.1038/nrn.2016.44.

[18] HH Mørch, 'Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism?' (2019) 84(5) *Erkenntnis* 1065–1085 https://doi.org/10.1007/s10670-018-9995-6.

2. *Privacy of Data and Information: Ownership, Authorship, Interest, and Responsibility*

Before delving into the current debate around mental privacy, let me provide a few propaedeutical thoughts on the terminology and conceptual foundations of privacy and how it is understood in the context of data and information processing. Etymologically, 'privacy' originates from the Latin term *privatus* which means 'withdrawn from public life'.[19] An important historical usage context for the concept of 'privacy' was in the military and warfare domain, for example in the notion of 'privateers', that is, a person or ship that privately participated in an armed naval conflict under official commission of war (distinguishing privateering from outlawed activities such as piracy).[20] The term and concept has a rich history in jurisprudence and the law. Lacking the space to retrace all ramifications of the legal-philosophical understandings of privacy, one notion that seems relevant for our context here – and that sets privacy apart from the related notion of seclusion and secrecy[21] – is that privacy ultimately concerns a person's 'autonomy within society'.[22] In the current age of digital information technology, this autonomy extends into the realm of the informational—in other words, the 'infosphere' as elucidated by *Luciano Floridi*[23]—which is reflected by an increasing number of ethical and legal analyses of 'informational privacy' and the metamorphosis of persons into 'data subjects' and digital service providers into 'data controllers' in the digital realm.[24] In this context, it may be worthwhile to remind us that data and information (and knowledge for that matter), though intricately intertwined, are not interchangeable notions. Whereas data are 'numbers and words without relationships', information are 'numbers and words with relationships' and knowledge refers to inferences gleaned from information.[25] This distinction is important for the development and application of granular and context-sensitive legal and policy instruments for protecting a person's privacy.[26]

For contexts in which questions around the protection of (and threats to) data or informational privacy are originating from the creation, movement, storage and analysis of digital data, it would seem appropriate to conceptualize 'informational privacy' as: autonomy of persons over the collection, access and use of data and information about themselves. Related to these questions, this expanding discussion has made the question of data (and information) ownership

[19] TF Hoad, 'Private' in TF Hoad (ed), *The Concise Oxford Dictionary of English Etymology* (2003) www.oxfordreference.com/view/10.1093/acref/9780192830982.001.0001/acref-9780192830982-e-11928.

[20] Another legacy in the military domain is the rank of private, i.e. soldiers of the lowest military rank.

[21] See e.g. the usage definition from Merriam Webster, 'Privacy' (*Merriam Webster Dictionary*) www.merriam-webster.com/dictionary/privacy.

[22] J Hirshleifer, 'Privacy: Its Origin, Function, and Future' (1980) 9(4) *The Journal of Legal Studies* 649–664.

[23] L Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality* (2014).

[24] AD Vanberg, 'Informational Privacy Post GDPR: End of the Road or the Start of a Long Journey?' (2021) 25(1) *The International Journal of Human Rights* 52–78 https://doi.org/10.1080/13642987.2020.1789109 (hereafter Vanberg, 'Informational Privacy Post GDPR'); TW Kim and BR Routledge, 'Informational Privacy, A Right to Explanation, and Interpretable AI' in IEEE (ed), *2018 IEEE Symposium on Privacy-Aware Computing (PAC)* (2018) 64–74 https://doi.org/10.1109/PAC.2018.00013; AD Moore, 'Toward Informational Privacy Rights 2007 Editor's Symposium' (2007) 44(4) *San Diego Law Review* 809–846; L Floridi, 'Four Challenges for a Theory of Informational Privacy' (2006) 8(3) *Ethics and Information Technology* 109–119 https://doi.org/10.1007/s10676-006-9121-3 (hereafter Floridi, 'Four Challenges for a Theory of Informational Privacy').

[25] J Pohl, 'Transition From Data to Information' in *Collaborative Agent Design Research Center Technical Report - RESU72* (2001) 1–8.

[26] Depending on the context, a very different granularity of privacy protection might be necessary. Consider, for example, the difference between collecting only one specific type of biometric data (without other contextual data) vs. collecting multimodal personal data to glean health-related information in a consumer technology context, which would require different granularity of data and information protection.

a central aspect of ethical and legal scholarship and policy debates.[27] In a legal context, the protection of data or informational privacy are relevant, *inter alia*, in trade law (e.g. confidential trade secrets), copyright law, health law and many other legal areas. Importantly, however, individuals do not have property rights regarding their personal information, e.g. information about their body, health and disease in medical records.[28] Separate from the question of ownership of personal information is the question of authorship, in other words, who can be regarded as the creator of specific data and information about a person.[29] But, even in contexts in which persons are neither the author/creator nor the owner of data and information about themselves, they nevertheless have legitimate interests in protecting this information from being misused to their disadvantage, and therefore legitimate interest, and derived thereof, right, to keep it private. This right to informational privacy is now a fundamental tenet in consumer protection laws as well as data protection and privacy laws, for example the European Union's (EU) General Data Protection Regulation (GDPR).[30]

Finally, these questions of ownership, authorship, and interests in personal data and information – and the legal mechanisms for protecting the right to informational privacy – of course also raise the questions of responsibility for and stewardship of personal data and information to protect them from unwarranted access and from misuse. Typically, many different participants and stakeholders are involved in the creation, administration, distribution, and use of personal data and information (i.e. the creator(s)/author(s), owner(s), persons with legitimate and vested interests). Under many circumstances, this creates a problem of ascribing responsibility for data stewardship – a diffusion of responsibility. This may be further complicated by the fact that the creator of a particular set of personal information, the owner (and the person to whom these data and information pertain), may reside in different jurisdictions and may therefore be accountable to different data protection and privacy laws.

3. *Mental Privacy: Protecting Data and Information about the Human Brain and Associated Mental Phenomena*

In the debate around 'neurorights' the term mental privacy has established itself to refer to the 'mental realm' outlined above. However, from a materialist, neurodeterministic position, it would not make much sense to give mental phenomena special juridical protection if we neither have ways to measure these phenomena nor a model of causal mechanisms to give an account of how they arise. For the law, however, such a strict mechanistic interpretation of

[27] A Ballantyne, 'How Should We Think about Clinical Data Ownership?' (2020) 46(5) *Journal of Medical Ethics* 289–294 https://doi.org/10.1136/medethics-2018-105340; P Hummel, M Braun and P Dabrock, 'Own Data? Ethical Reflections on Data Ownership' (2020) *Philosophy & Technology* 1-28 https://doi.org/10.1007/s13347-020-00404-9; M Mirchev, I Mircheva and A Kerekovska, 'The Academic Viewpoint on Patient Data Ownership in the Context of Big Data: Scoping Review' (2020) 22(8) *Journal of Medical Internet Research* https://doi.org/10.2196/22214; N Duch-Brown, B Martens and F Mueller-Langer, 'The Economics of Ownership, Access and Trade in Digital Data' (*SSRN*, 17 February 2017), https://doi.org/10.2139/ssrn.2914144.

[28] Canada Supreme Court, *McInerney v MacDonald* (11 June 1992) 93 Dominion Law Reports 415–31.

[29] JC Wallis and CL Borgman, 'Who Is Responsible for Data? An Exploratory Study of Data Authorship, Ownership, and Responsibility' (2011) 48(1) Proceedings of the American Society for Information Science and Technology 1–10 https://doi.org/10.1002/meet.2011.14504801188.

[30] Vanberg, 'Informational Privacy Post GDPR' (n 24); FT Beke, F Eggers, and PC Verhoef, 'Consumer Informational Privacy: Current Knowledge and Research Directions' (2018) 11(1) *Foundations and Trends(R) in Marketing* 1–71; HT Tavani, 'Informational Privacy: Concepts, Theories, and Controversies' in KH Himma and HT Tavani (eds), *The Handbook of Information and Computer Ethics* (2008) 131–64 https://doi.org/10.1002/9780470281819.ch6; Floridi, 'Four Challenges for a Theory of Informational Privacy' (n 24).

mental mechanisms might not be required to ensure adequate protections. Consider, for example, that crimes with large immaterial components such as 'hate speech' or 'perjury' also contain a large component of internal processes that might remain hidden from the eye of the law. In hate speech, for instance, both the level of internal motivation of the perpetrator as well as the level of internal processes of psychological injury in the injured party do not need to be objectivated in order to establish whether or not a punishable crime was committed.

The precise understanding and interpretation of mental privacy also differs substantially across literatures, contexts, and debates. In legal philosophy, for instance, mental privacy is mainly discussed in the context of foundational questions and justifications in criminal justice such as the concept of *mens rea* (the 'guilty mind'),[31] freedom of the will, the feasibility of lie detection, and other 'neurolaw' issues.[32] In neuroethics, mental privacy is often invoked in discussions around brain data governance and regulation as well as in reference to 'neurorights': the question of whether the protection of mental privacy is (or shall become) a part of human rights frameworks and legislation.[33] The discussion here shall be concerned with the latter context.

III. MENTAL INTEGRITY THROUGH THE LENS OF VULNERABILITY ETHICS

Mental integrity, much like the term mental privacy, has an evocative appeal which allows for an intuitive and immediate approximate understanding: to protect the intactness and inviolacy of brain structure and functions (and the associated mental experiences).

Like mental privacy, however, mental integrity is currently still lacking a broadly accepted definition across philosophy, ethics, cognitive science, and neuroscience.[34] Most operational descriptions refer to the idea that the structure and function of the human brain and the corresponding mental experiences allow for an integrated mental experience for an individual and that external interference with this integrated experience requires a reasonable justification (such as medication for disturbed states of mind in psychosis, for example) to be morally (and legally) acceptable. The problem that the nature of subjective mental experience, phenomenal consciousness, is inaccessible both internally (as the subject can only describe the qualitative aspects of the experience itself, but not the mechanics of its composite nature) and externally, also affects the way in which we conceptualize the notion of an integrated mind. As an individual – the indivisible person in the literal sense – we mostly experience the world in a more or less unified way, even though separate parallel perceptual, cognitive, and emotive processes have to be integrated in a complex manner to allow for this holistic experience. When being asked, for example, by a curious experimental psychologist or cognitive scientist, to describe the nature of our experience, for example seeing a red apple on a table, we can identify qualitative characteristics of the apple: its shape, texture, colour, and perhaps smell. Yet, we have no shared terminology to describe the quality of our inner experience of seeing the apple – outside of associating particular thoughts, memories, or emotions with this instance of an apple or apples in general.

[31] P Kellmeyer, 'Ethical and Legal Implications of the Methodological Crisis in Neuroimaging' (2017) 26(4) *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees* 530–554 https://doi.org/10.1017/S096318011700007X.

[32] G Meynen, 'Neurolaw: Neuroscience, Ethics, and Law. Review Essay' (2014) 17(4) *Ethical Theory and Moral Practice* 819–829 http://www.jstor.org/stable/24478606; TM Spranger, 'Neurosciences and the Law: An Introduction' in TM Spranger (ed), *International Neurolaw* (2012) 1–10 https://doi.org/10.1007/978-3-642-21541-4_1.

[33] Yuste and others, 'Four Ethical Priorities for Neurotechnologies and AI' (n 10); Ienca and Andorno, 'Towards New Human Rights in the Age of Neuroscience and Neurotechnology' (n 10); Kellmeyer, 'Big Brain Data' (n 9).

[34] A Lavazza, 'Freedom of Thought and Mental Integrity: The Moral Requirements for Any Neural Prosthesis' (*Froniters in Neuroscience*, 19 February 2018) 12 https://doi.org/10.3389/fnins.2018.00082.

Put in another way: We all know intuitively what a unified or integrated experience of seeing an apple is like but we cannot explain it in such a way that the descriptions necessarily evoke the same experience(s) in others. To better understand what an integrated experience is like, we might also consider what a disintegrated, disunified, or fragmented experience is like. In certain dream-like states, pathogenic states like psychosis or under the influence of psychoactive substances, an experience can disintegrate into certain constitutive components (e.g. perceiving the shape and colour of the apple separately, yet, simultaneously) or perceptions can be qualitatively altered in countless ways (consider, for instance, the phenomenon of synaesthesia, 'seeing' tones or 'hearing' colours). This demonstrated potential for the composite nature of mental experiences suggests that it is not inconceivable that we might find more targeted and precise ways to influence the qualitative nature (and perhaps content) of our mental experiences, for example, through precision drugs or neurotechnological interventions.[35] Emerging techniques such as optogenetics, for instance, have already been demonstrated to be able to 'incept' false memories into a research animal's brain.[36] But our mental integrity can also be compromised by non-neurotechnological interventions of course. Consider approaches from (behavioral) psychology such as nudging or subliminal priming (and related techniques)[37] that can influence decision making and choice (and have downstream effects on the experiences associated with these decisions and choices) or more overt psychological interventions such as psychotherapy or the broad – and lately much questioned (in the context of the replication crisis in psychology[38]) – field of positive psychology, for example mindfulness,[39] meditation, and related approaches.

Direct neurotechnologically mediated interventions into the brain intuitively raise health and safety concerns, for example concerning potential adverse effects on mental experience and therefore mental integrity. While such safety concerns are surely reasonable given the direct physical nature of the brain intervention, there is, however, to date no evidence of serious adverse effects for commonly used extracranial electric or electromagnetic neurostimulation techniques such as transcranial direct-current stimulation (tDCS) or repetitive transcranial magnetic stimulation (rTMS).[40] In stark contrast, comparatively little attention has been paid until recently to the adverse effects of psychological interventions. Studies in the past few years have now demonstrated that seemingly benign interventions such as psychotherapy, mindfulness, or meditation can have discernible and sometimes serious adverse effects on mental health and well-being and thus on mental integrity.[41]

[35] F Germani and others, 'Engineering Minds? Ethical Considerations on Biotechnological Approaches to Mental Health, Well-Being, and Human Flourishing' (*Trends in Biotechnology*, 3 May 2021) https://doi.org/10.1016/j.tibtech.2021.04.007; P Kellmeyer, 'Neurophilosophical and Ethical Aspects of Virtual Reality Therapy in Neurology and Psychiatry' (2018) 27(4) *Cambridge Quarterly of Healthcare Ethics* 610–627 https://doi.org/10.1017/S0963180118000129.

[36] CK Kim, A Adhikari, and K Deisseroth, 'Integration of Optogenetics with Complementary Methodologies in Systems Neuroscience' (2017) 18(4) *Nature Reviews Neuroscience* 222–235 https://doi.org/10.1038/nrn.2017.15.

[37] C Janiszewski and RS Wyer, 'Content and Process Priming: A Review' (2014) 24(1) *Journal of Consumer Psychology* 96–118 https://doi.org/10.1016/j.jcps.2013.05.006; DM Hausman, 'Nudging and Other Ways of Steering Choices' (2018) 1 *Intereconomics* 17–20.

[38] Open Science Collaboration, 'Estimating the Reproducibility of Psychological Science' (2015) 349(6251) *Science* https://doi.org/10.1126/science.aac4716.

[39] JD Creswell, 'Mindfulness Interventions' (2017) 68(1) *Annual Review of Psychology* 491–516 https://doi.org/10.1146/annurev-psych-042716-051139.

[40] H Matsumoto and Y Ugawa, 'Adverse Events of TDCS and TACS: A Review' (2017) 2 *Clinical Neurophysiology Practice* 19–25 https://doi.org/10.1016/j.cnp.2016.12.003; F Fregni and A Pascual-Leone, 'Technology Insight: Noninvasive Brain Stimulation in Neurology—Perspectives on the Therapeutic Potential of RTMS and TDCS' (2007) 3(7) *Nature Clinical Practice Neurology* 383–393 https://doi.org/10.1038/ncpneuro0530.

[41] AWM Evers and others, 'Implications of Placebo and Nocebo Effects for Clinical Practice: Expert Consensus' (2018) 87(4) *Psychotherapy and Psychosomatics* 204–210 https://doi.org/10.1159/000490354; WB Britton and others, 'Defining

Another context in which there is intensive debate around the ethical aspects and societal impact of influencing mental experience and behavior concerns internet-based digital technologies, especially the issue of gamification[42] and other incentivizing forms of user engagement in 'social' media platforms or apps. Certain types of digital behavioral technologies[43] are specifically designed to tap into reward-based psychological and neurobiological mechanisms with the aim to maximize user engagement which drives the business model of many companies and developers in the data economy.[44] While these digital behavioral technologies (DBT) might be used in a healthcare provision context, for example to deliver digital mental health services,[45] the use of DBT apps in an uncontrolled environment, such as internet-based media and communication platforms raises concern about the long-term impact on mental integrity of users.

To summarize, the quality and content of our mental experience is multifaceted and the ability to successfully integrate different levels of mental experience into a holistic sense of self (as an important component of selfhood or personhood) – mental integrity – is an important prerequisite for mental health and well-being. There are several ways to interfere with mental integrity, through neurotechnologically mediated interventions as well as by many other means. The disruption of the integrated nature of our mental life can lead to severe psychological distress and potentially mental illness. Therefore, protecting our mental life from unwarranted and/or unconsented intervention seems like a justified ethical demand. The law offers many mechanisms for protection in that respect, both at the level of fundamental rights – for example in Article 3 – Right to integrity of the person of the EU Charter of Fundamental Rights[46] – as well as specific civil laws such as consumer protection laws and medical law.

IV. NEURORIGHTS: LEGAL INNOVATION OR NEW WINE IN LEAKY BOTTLES?

As we have seen in the preceding sections, there are ethically justifiable and scientifically informed reasons to claim that mental privacy and mental integrity are indeed aspects of our human existence ('anthropological goods' if you will) that are worthy of being protected by the law. In this section, I will therefore give an overview of recent developments in the legal and

and Measuring Meditation-Related Adverse Effects in Mindfulness-Based Programs' (*Clinical Psychological Science*, 18 May 2021) https://doi.org/10.1177/21677026211996340; M Farias and others, 'Adverse Events in Meditation Practices and Meditation-Based Therapies: A Systematic Review' (2020) 142(5) *Acta Psychiatrica Scandinavica* 374–393 https://doi.org/10.1111/acps.13225; D Lambert, NH van den Berg, and A Mendrek, 'Adverse Effects of Meditation: A Review of Observational, Experimental and Case Studies' (*Current Psychology*, 24 February 2021) https://doi.org/10.1007/s12144–021-01503-2.

[42] A Hoffmann, CA Christmann, and G Bleser, 'Gamification in Stress Management Apps: A Critical App Review' (2017) 5(2) *JMIR Serious Games* https://doi.org/10.2196/games.7216.

[43] L Herzog, P Kellmeyer, and V Wild, 'Digital Behavioral Technology, Vulnerability and Justice: An Integrated Approach' (*Review of Social Economy*, 30 June 2021) www.tandfonline.com/doi/full/10.1080/00346764.2021.1943755?scroll=top&needAccess=true (hereafter Herzog, Kellmeyer, and Wild, 'Digital Behavioral Technology, Vulnerability and Justice: An Integrated Approach').

[44] T Wu, *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (2017); AA Alhassan and others, 'The Relationship between Addiction to Smartphone Usage and Depression Among Adults: A Cross Sectional Study' (*BMC Psychiatry*, 25 May 2018) https://doi.org/10.1186/s12888-018-1745-4; DT Courtwright, *Age of Addiction: How Bad Habits Became Big Business* (2021); NM Petry and others, 'An International Consensus for Assessing Internet Gaming Disorder Using the New DSM-5 Approach' (2014) 109(9) *Addiction* 1399–1406 https://doi.org/10.1111/add.12457.

[45] VW Sze Cheng and others, 'Gamification in Apps and Technologies for Improving Mental Health and Well-Being: Systematic Review' (2019) 6(6) *JMIR Mental Health* https://doi.org/10.2196/13717.

[46] EU: Council of the European Union, 'Charter of Fundamental Rights of the European Union', C 303/1 2007/C 303/01 § (2007).

policy domain regarding the implementation of such 'neurorights'.[47] First, I will describe the current debate around the legal foundations and scope of neurorights, then I will propose some conceptual additions to the notion of neurorights and, third, propose a pragmatic and human rights–based approach for making neurorights actionable.

1. *The Current Debate on the Conceptual and Normative Foundations and the Legal Scope of Neurorights*

For a few years now, the debate around the legal foundations and precise scope of neurorights has been steadily growing. From a bird's eye perspective, it seems fair to say that two main positions are dominating the current scholarly discourse: rights conservatism and rights innovationism/reformism. Scholars that argue from a rights conservatism position make the case that the existing set of fundamental rights, as enshrined for example in the Universal Declaration of Human Rights (UDHR) (but also in many constitutional legal frameworks in different states and specific jurisdictions), provides enough coverage to protect the anthropological goods of mental privacy and mental integrity.[48] Scholars that are arguing from the position of rights innovationism or reformism emphasize that there is something qualitatively special and new about the ways in which emerging neurotechnologies (and other methods, see above) (may) allow for unprecedented access to a person's mental experience or (may) interfere with their mental integrity, and that, therefore, either new fundamental rights are necessary (legal innovation) or existing fundamental rights should be amended or expanded (legal reformism).[49] Common to both positions is the acknowledgment that the privacy and integrity of mental experience are indeed aspects of human existence that should be protected by the law; the differences in terms of how such mental protection could be implemented, however, have vastly different implications in terms of the consequences for national and international law. Whereas the legal conservatist would have to do the work to show precisely how national, international, and supranational legal frameworks could be effectively applied to protect mental privacy and integrity in specific contexts, the reformist position implies changes in the legal landscape that would have seismic and far-reaching consequences for many areas of the law, national and international policymaking as well as consumer protection and regulatory affairs. From a pragmatic point of view, two major problems immediately present themselves regarding the addition of new fundamental rights that refer to the protection of mental experience to the catalogue of human rights. The first problem concerns the potential for unintended consequences of introducing such novel rights. It is a well-known problem, both in moral philosophy and legal philosophy, that moral and legal goods – especially if they are not conceptually dependent on each other – can (and often do) exist in conflict with each other which, in applied moral philosophy gives rise to classical dilemma situations for example. Therefore, introducing new fundamental rights might serve

[47] As I am not a legal scholar, this section provides an outside view, informed by my understanding of the neuroscientific facts and ethical discussions, of the current debate at the intersection of neurolaw and neuroethics on the relevance of fundamental rights, particularly international human rights, for protecting mental privacy and mental integrity. In the scholarly debate, this set of issues are usually referred to as 'neurorights' and I will therefore use this term here too.

[48] S Ligthart and others, 'Forensic Brain-Reading and Mental Privacy in European Human Rights Law: Foundations and Challenges' (*Neuroethics*, 20 June 2020) https://doi.org/10.1007/s12152-020-09438-4; C Bublitz, 'Cognitive Liberty or the International Human Right to Freedom of Thought' in J Clausen and N Levy (eds), *Handbook of Neuroethics* (2015) 1309–1333 https://doi.org/10.1007/978-94-007-4707-4_166.

[49] Yuste and others, 'Four Ethical Priorities for Neurotechnologies and AI' (n 10); Ienca and Andorno, 'Towards New Human Rights in the Age of Neuroscience and Neurotechnology' (n 10).

the purpose of protecting a specific anthropological good, such as mental privacy, in a granular way, but at the same time it increases the complexity of balancing different fundamental rights and therefore also the potential for moral and/or legal dilemmata situations. Another often voiced criticism is the perceived problem of rights inflation, in other words, the notion that the juridification (German: '*Verrechtlichung*') of ethical norms leads to an inflation of fundamental rights – and thus rights-based narratives and juridical claims – that undermine the ability of the polity to effectively address systemic social and other structural injustices.[50]

From my point of view, the current state of this debate suffers from the following two major problems: firstly, an insufficient conceptual specification of mental privacy and mental integrity and, secondly, a lack of transdisciplinary collaborative discourses and proposals for translating the ethical demands that are framed as neurorights into actionable frameworks for responsible and effective governance of neurotechnologies. In the following sections, I address both concerns by suggesting some conceptual additions to the academic framing and discourse around neurorights and proposing a strategy for making neurorights actionable.

2. *New Conceptual Aspects: Mental Privacy and Mental Integrity As Anthropological Goods*

The variability of operational descriptions of mental privacy and mental integrity in the literature shows that both notions are still 'under construction' from a conceptual perspective. As important as this ongoing conceptual work is in refining these ideas and for making them accessible to a wide scholarly audience, I would propose here that understanding them mainly as relevant anthropological goods[51] – rather than mainly philosophical or legal concepts – could help to theorize and discuss about mental privacy and mental integrity across disciplinary divides. However, the anthropological goods of mental privacy and mental integrity are conceptually underspecified in the following sense.

First, no clear account is given in the literature of what typical, if not the best approximate, correlates of mental experience (as the substrate of mental privacy) are. Some authors suggest that neurodata or brain data are – or might well become (with advances in neuroscience) – the most direct correlate of mental experience and that, therefore, brain data (and information gleaned from these data) should be considered a noteworthy and special category of personal data.[52] It could be argued that, in addition to brain data, many different kinds of contextual data (e.g. from smartphones, wearables, digital media and other contexts) allow for similar levels of diagnostic or predictive modelling and inferences on the quality and content of a person's

[50] D Clément, 'Human Rights or Social Justice? The Problem of Rights Inflation' (2018) 22(2) *The International Journal of Human Rights* 155–169 https://doi.org/10.1080/13642987.2017.1349245. Though there are also important objections to these lines of arguments: JT Theilen, 'The Inflation of Human Rights: A Deconstruction' (2021) *Leiden Journal of International Law* 1–24 https://doi.org/10.1017/S0922156521000297.

[51] An anthropological good, in my usage here, refers to a key foundational dimension of human existence that, throughout history and across cultures, is connected to strong human interests and preferences. Examples would be the interest in and preference for being alive, for having shelter, freedom, food, and so forth. In this understanding, anthropological goods antecede and often are the basis for normative demands, such as ethical claims and rights claims. As a pre-theoretical notion, they are also related to the more developed notion of 'capabilities' [M Nussbaum, 'Capabilities and Social Justice' (2002) 4(2) *International Studies Review* 123–135 https://doi.org/10.1111/1521-9488.00258] insofar as capabilities give a philosophically comprehensive account of how dimensions of human existence relate to fundamental rights.

[52] Kellmeyer, 'Big Brain Data' (n 9); Sara Goering and others, 'Recommendations for Responsible Development and Application of Neurotechnologies' (2021) *Neuroethics* https://doi.org/10.1007/s12152-021-09468-6.

mental experience.[53] What is lacking, however, is a critical discussion of what the right level for protecting a person's mental privacy is: the level of data protection (data privacy); protecting the information/content that can be extracted from these data (informational privacy); or both; or whether we should also address the question of how and to what ends mental data/information are being used? As discussed above, I would suggest that a very important and legitimate dimension for ethical concerns is also the question of whether and to what extent any kind of neurotechnology or neurodecoding approach has a negative impact on enabling a person to exercise their legitimate interest in their own mental data and information. To be able to respect a person's interest in data and information on their mental states, however, we would need ethically viable means of disclosing these interests to a third party in ways that do not themselves create additional problems of privacy protection, in other words to avoid a self-perpetuating privacy protection problem. At the level of data and information protection, one strategy could be to establish trustworthy technological means (such as blockchain technology, differential privacy, homomorphic encryption, and other techniques[54]) and/or institutions – data fiduciaries – for handling any data of a person that might allow for inferences on mental experience.

Second, the demand for protecting mental integrity is undermined by the problem that we do not have a consensual conceptual understanding of key notions such as agency, autonomy, and the self. Take the example of psychedelic recreational drugs, as an example for an outside interference with mental integrity. We have ample evidence from psychological and psychiatric research that suggests that certain types of recreational psychedelic drugs, such as LSD or Psylocibin, have discernible effects on mental experiences associated with personal identity and self-experience, variously called, for example, 'ego dissolution'[55] or 'boundlessness'.[56] However, most systematic research studying these effects, say in experimental psychology or psychiatry, is not predicated on a universal understanding or model of human self-experience, personal identity, and related notions. As even any preliminary engagement with conceptual models of personal identity or 'the' self in psychology, cognitive science, and philosophy will quickly reveal, there are indeed many different competing, often conceptually non-overlapping or incommensurable models available: ranging from constructivist ideas of a 'narrative self', to embodiment-related (or more generally 4E-cognition-related) notions of an 'embodied self' or 'active self', to more socially inspired notions such as the 'relational self' or 'social self'.[57]

[53] Herzog, Kellmeyer, and Wild, 'Digital Behavioral Technology, Vulnerability and Justice: An Integrated Approach' (n 42); KV Kreitmair, MK Cho, and DC Magnus, 'Consent and Engagement, Security, and Authentic Living Using Wearable and Mobile Health Technology' (2017) 35(7) *Nature Biotechnology* 617–620 https://doi.org/10.1038/nbt.3887; N Minielly, V Hrincu, and J Illes, 'A View on Incidental Findings and Adverse Events Associated with Neurowearables in the Consumer Marketplace' in I Bárd and E Hildt (eds), *Developments in Neuroethics and Bioethics*, *vol.* 3 (2020) 267–277 https://doi.org/10.1016/bs.dnb.2020.03.010.

[54] V Jaiman and V Urovi, 'A Consent Model for Blockchain-Based Health Data Sharing Platforms' in *IEEE Access* 8 (2020) 143734–143745 https://doi.org/10.1109/ACCESS.2020.3014565; A Khedr and G Gulak, 'SecureMed: Secure Medical Computation Using GPU-Accelerated Homomorphic Encryption Scheme' (2018) 22(2) *IEEE Journal of Biomedical and Health Informatics* 597–606 https://doi.org/10.1109/JBHI.2017.2657458; MU Hassan, MH Rehmani, and J Chen, 'Differential Privacy Techniques for Cyber Physical Systems: A Survey' (2020) 22(1) *IEEE Communications Surveys Tutorials* 746–789 https://doi.org/10.1109/COMST.2019.2944748.

[55] C Letheby and P Gerrans, 'Self Unbound: Ego Dissolution in Psychedelic Experience' (2017) 1 *Neuroscience of Consciousness* https://doi.org/10.1093/nc/nix016.

[56] FX Vollenweider and KH Preller, 'Psychedelic Drugs: Neurobiology and Potential for Treatment of Psychiatric Disorders' (2020) 21(11) *Nature Reviews Neuroscience* 611–624 https://doi.org/10.1038/s41583-020-0367-2.

[57] PT Durbin, 'Brain Research and the Social Self in a Technological Culture' (2017) 32(2) *AI & SOCIETY* 253–260 https://doi.org/10.1007/s00146-015-0609-4; S Gallagher, 'A Pattern Theory of Self' (2013) 7 *Frontiers in Human Neuroscience* https://doi.org/10.3389/fnhum.2013.00443; T Fuchs, *The Embodied Self: Dimensions, Coherence, and Disorders* (2010); D Parfit, 'Personal Identity' (1971) 80(1) *The Philosophical Review* 3–27.

Consequently, any interpretation, let alone systematic understanding, of how certain interventions might or might not affect mental integrity – here represented by the dimension of self-experience and personal identity – will heavily depend on the conceptual model of mental experience that one has. This rather obvious point about the inevitable interdependencies between theory-driven modelling and data-driven inferences and interpretation has important consequences for the ethical demands and rights-claims that characterize the debate on the neurorights. First, this should lead to the demand and recommendation that any empirical research that investigates the relationship between physical (for instance via neurotechnologies or drugs) or psychological interventions (for example through behavioural psychology, such as nudging) and mental experience should make their underlying model of self-experience and personal identity explicit and specify it in a conceptually rigorous manner. Second, transdisciplinary research on the conceptual foundations of mental (self-)experience, involving philosophers, cognitive scientists, psychologists, neuroscientists, and clinicians should be encouraged to arrive at more widely accepted working models that can then be tested empirically.

3. *Making Neurorights Actionable and Justiciable: A Human Rights–Based Approach*

Irrespective of whether new fundamental rights will ultimately be deemed necessary or whether existing fundamental rights will prove sufficient to protect the anthropological goods mental privacy and mental integrity, regulation and governance of complex emerging sciences and technologies, such as AI-based neurotechnology, is a daunting challenge. If one would agree that reasonable demands for any governance regime that allows innovation of emerging technologies in a responsible manner include that the regime is context-sensitive, adaptive, anticipatory, effective, agile, and at the right level of ethical and legal granularity, then the scattered and inhomogeneous landscape of national and international regulatory and legal frameworks and instruments presents a particularly complex problem of technology governance.[58]

Apart from the conceptual issues discussed here that need to be further clarified to elucidate the basis for specific ethical/normative demands for protecting mental privacy and mental integrity, another important step for making neurorights actionable is finding the right levels of governance and regulation and appropriate (and proportional) granularities of legal frameworks. So far, no multi-level approach to legal protection of mental privacy and mental integrity is available. Instead, we find various proposals and initiatives at different levels: at the level of ethical self-regulation and self-governance; represented for example by ethical codes of conduct in the context of neuroscience research[59] or in the private sector around AI governance;[60] at the level of national policy, regulatory, and legislative initiatives (e.g. in Chile);[61] at the level of supranational policies and treaties, represented, for example, by the intergovernmental report on

[58] More generally the complexity of the legal landscape and political processes creates the well-known 'pacing problem' in governing and regulating technological innovations, also referred to as the 'Collingridge Dilemma', *cf.* for example: A Genus and A Stirling, 'Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation' (2018) 47(1) *Research Policy* 61–69 https://doi.org/10.1016/j.respol.2017.09.012.

[59] Exemplified by the Ethics Policy of the Society for Neuroscience, the largest professional body representing neuroscience researchers: SfN, 'Professional Conduct' (*SfN*) https://www.sfn.org/about/professional-conduct.

[60] Consider for example: *Partnership on AI* www.partnershiponai.org/.

[61] L Dayton, 'Call for Human Rights Protections on Emerging Brain-Computer Interface Technologies' (*Nature Index*, 16 March 2021) https://www.natureindex.com/news-blog/human-rights-protections-artificial-intelligence-neurorights-brain-computer-interface.

responsible innovation in neurotechnology of the Organization for Economic Co-operation and Development (OECD) from 2019[62].

Taking these complex problems into account, I would advocate for a pragmatic, human rights–based approach to regulating and governing AI-based neurotechnologies and for protecting mental privacy and mental integrity as anthropological goods. This approach is predicated on the assumption that existing fundamental rights, as enshrined in the UDHR and many national constitutional laws, such as the right to freedom of thought,[63] provide sufficient normative foundations. On top of these foundations, however, a multi-level governance approach is required that provides context-sensitive and adaptive regulatory, legal, and political solutions (at the right level of granularity) for protecting humans from potential threats to mental privacy and mental integrity, such as in the context of hitherto un- or underregulated consumer neurotechnologies. Such a complex web of legal and governance tools will likely include bottom-up instruments, such as ethical self-regulation, but also laws (constitutional laws, but also consumer protection laws and other civil laws) and regulations (data protection regulations and consumer regulations) at the national level and supranational level, as well as soft-law instruments at the supranational level (such as the OECD framework for responsible innovation of neurotechnology, or widely adopted ethics declarations from specialized agencies of the United Nations (UN), such as UN Educational, Scientific and Cultural Organization (UNESCO) or World Health Organization (WHO)).

But making any fundamental right actionable (and justiciable) at all levels of societies and international communities requires a legally binding and ethically weighty framework to resolve current, complex, and controversial issues in science, society, and science policy. Therefore, conceptualizing neurorights as a scientifically grounded and normatively oriented bundle of fundamental rights (and applied legal and political translational mechanisms) may have substantial inspirational and instrumental value for ensuring that the innovation potential of neurotechnologies, especially AI-based approaches, can be leveraged for applications that promote human health, well-being, and flourishing.

V. SUMMARY AND CONCLUSIONS

In summary, neurorights have become an important subject for scholarly debate, driven partly by innovation in AI-based decoding of neural activity, and as a result different positions are emerging in the discussion around the legal status of brain data and the legal approach to protecting the brain and mental content from unwarranted access and interference.

I have argued that mental privacy and mental integrity could be understood as important anthropological goods that need to be protected from unwarranted and undue interference, for example, by means of neurotechnology, particularly AI-based neurotechnology.

In the debate on the question of how neurorights relate to existing national and supranational legal frameworks, especially to human rights, three distinct positions are emerging: (a) a rights conservatism position, in which scholars argue that existing fundamental rights (e.g. constitutional rights at the national level and human rights at the supranational level) provide adequate protection to mental privacy and mental integrity; (b) a reformist, innovationist position, in which scholars argue that existing legal frameworks are not sufficient to protect

[62] OECD Legal Documents, 'Recommendation of the Council on Responsible Innovation in Neurotechnology' https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0457.

[63] Article 18, UDHR.

the brain and mental content of individuals under envisioned near-future scenarios of AI-based brain decoding through neurotechnologies and, therefore, reforms of existing frameworks – such as constitutional laws or even the Universal Declaration of Human Rights – are required; and (c) a human rights–based approach, that acknowledges that the law (in most national jurisdictions as well as internationally) provides sufficient legal instruments but that its scattered nature – across jurisdictions as well as different areas and levels of the law (such as consumer protection laws, constitutional rights, etc.) – requires an approach that makes neurorights actionable and justiciable, for example by connecting fundamental rights to specific applied laws (e.g. in consumer protection laws).

The latter position – which in the policy domain would translate into a multi-level governance approach – has the advantage that it does not argue from entrenched positions with little room for consilience but provides deliberative space in which agreements, treaties, soft law declarations, and similar instruments for supra- and transnational harmonization can thrive.

25

AI-Supported Brain–Computer Interfaces and the Emergence of 'Cyberbilities'

Boris Essmann and Oliver Mueller

I. INTRODUCTION

Recent advances in brain–computer interfacing (BCI) technology hold out the prospect of technological intervention into the basis of human agency to supplement and restore functioning in agency-limited individuals and even augmenting and enhancing capacities for natural agency. By increasingly using Artificial Intelligence (AI), for example machine learning methods, a new generation of brain–computer interfaces aims to advance technological possibilities to intervene into agentive capacities even more, creating new forms of human–machine interaction in the process. This trend further accentuates concerns about the impact of neurotechnology on human agency, not only regarding far-reaching visions like the media-effective propositions by *Elon Musk* (Neuralink) but also with respect to current developments in medicine. Because these developments could be understood as (worrisome) 'fusions' of human, machinic, and software agency we investigate neurotechnology and AI-assisted brain–computer interfaces by directly focusing on agentive dimensions and potential changes of agency in these types of interactions. By providing a philosophical discussion of these topics we aim to capture the broad impact of this technology on our future and contribute valuable perspectives on its ethically and socially relevant dimensions. Although we adopt a philosophical approach, we do not restrict ourselves to a single disciplinary perspective, such as an exclusively ethical or neuroscience-oriented analysis. Given the potential to fundamentally reshape our individual and collective lives, the combination of neurotechnology and AI-technology may well create challenges that exceed disciplinary boundaries and which, therefore, cannot be met by a single discipline.

Our contribution to discussing the 'fusion' of human and artificial agency is the introduction of two neologisms – cyberbilities and hybrid agency – which we understand as concepts that integrate a range of disciplinary perspectives on this phenomenon. At a fundamental level, the concept loosely draws on *Amartya Sen*'s and *Martha Nussbaum*'s capabilities approach, but retools the notion of capabilities to analyze intricate human–machine interactions. We specifically adopt the normative core of capabilities – the ethical value of well-being opportunities – as a conceptual tool to evaluate risks and benefits of AI-supported brain–computer interfaces. However, like capabilities, cyberbilities presuppose a concept of human agency. Therefore, devising this concept requires a clarification of the underlying understanding of agency. Furthermore, because cyberbilities involve agency that is assisted by neurotechnology, we will also include an analysis of the various interactions between human and non-human elements involved.

This chapter is divided into three main sections. In the first section, we present conceptual expositions of the terms capabilities, agency, and human–machine interaction which serve both as an illustration of the complex nature of BCI technology and some necessary background to motivate the following line of argument.[1] This section is not intended to exhaust the topic from a specific (e.g., ethical or neuroscientific) perspective, but rather to amalgamate three very different but – as we maintain – complementary approaches. Specifically, we draw on the work of capability theorists such as *Sen* and *Nussbaum*.[2] Also, since neurotechnology affects human agency on various levels, we discuss the notions of agency and human–machine interaction from the perspectives of neuroscience, philosophical action theory, and a sociological framework.[3] In the next section, we introduce the above-mentioned novel concepts of hybrid agency and cyberbilities which combine our preceding line of argument and denote new forms of agency resulting from 'agentive' technologies.[4] A cyberbility is a type of capability, in other words, it is a normative concept designed to gauge the various ways in which neurotechnology can lead to achievements of (or want of) well-being and contribute to (or detract from) human flourishing. In the last section, we propose a list of cyberbilities that illustrates ways in which neurotechnology can lead to well-being gains (or losses) and explores the personal, social, and political ramifications of neurotechnologically assisted (or, in our terms, hybrid) agency.[5] However, this list of cyberbilities should not be understood as a conclusive result of the preceding conceptual work, but rather as a tentative and incomplete catalogue of core claims and requirements that reflect how new kinds of technologies challenge our established understanding of agency and human–machine interaction. In this sense, we see the list of cyberbilities not as a completed ethical evaluation, but as a foray into mapping tentative points of normative orientation.[6] And finally, we want to discuss a potential objection regarding our approach.[7]

II. FROM CAPABILITIES TO CYBERBILITIES

Let's start by anticipating our definition of cyberbilities: Cyberbilities are capabilities that originate from hybrid agency (i.e. human–machine interactions), in which agency is distributed across human and neurotechnological elements. As we will lay out in the following sections, this definition emphasizes that cyberbilities are embedded not only in personal aspects of agency, but also in a social environment that is shaped by the 'logic' of the respective technology and the institutions that deploy it (i.e. the 'technological condition').

In order to provide the necessary background for the notion of cyberbilities, we shall proceed in three steps. Firstly, we will briefly unfold in which way we retool the capabilities approach for our own purposes. Secondly, we argue that we need to revisit the concept of agency concerning its use in neuroscience and philosophy if we want to reliably describe the complex interactions between human and artificial elements, especially in the context of brain–computer interfaces. Lastly, we will draw on the notion of distributed agency introduced by sociologist *Werner Rammert*[8] to illuminate how technology affects agency and, consequently, human–machine

[1] See Section II.
[2] See Sub-section II 1.
[3] See Sub-section II 2.
[4] See Section III.
[5] See Sub-section IV 1.
[6] See Sub-section IV 2.
[7] See Section V.
[8] W Rammert, 'Where the Action Is: Distributed Agency between Humans, Machines, and Programs' in U Seifert, JH Kim, and A Moore (eds) *Paradoxes of Interactivity* (2008) (hereafter Rammert, 'Distributed Agency').

interactions. All three steps serve to review current disciplinary views on the topics at hand and prepare our proposal of an extended and integrated perspective in Section III.

1. Capabilities

The capabilities approach, first introduced by *Sen*[9] and extended by *Nussbaum*[10], is a theoretical framework used in a number of fields to evaluate the well-being of individuals in relation to their social, political, and psychological circumstances. To capability theorists, each person can be described (and thus compared) in terms of their 'capabilities' to achieve and maintain well-being, and any restrictions of those capabilities are subject to ethical scrutiny. As a philosophical term, well-being does not mean, for example, happiness, wealth, or absence of negative emotions or circumstances. Rather, well-being is meant to encompass how well a person's life is going overall, not just in relation to available means to lead a comfortable life or to achieve temporary positive emotional states, but concerning that a person is understood as an end when we focus on the opportunities to lead a good life that are available to each person.[11]

There is a long history of debate on the capabilities approach, and *Sen* and *Nussbaum* themselves delivered further refinements of the approach. We are aware of the fact that there are a number of controversies and open questions, for example, that *Sen*'s account is overly individualistic[12], or regarding certain essentialist traits[13] of *Nussbaum*'s version of the capabilities approach. However, due to the explorative purpose of this paper, we do not want to engage in further discussions of these aspects. Rather, we draw on *Sen*'s and *Nussbaum*'s theories in a pragmatic way, adopting some of their core elements in order to develop a basis for our tentative list of cyberbilities, which we see as a conceptual means not only to grasp novel kinds of agency in the upcoming age of human–machine fusions but also to propose a perspective that could help to evaluate these human–machine mergers as well.

But what are capabilities? Loosely following *Sen*, a capability describes what a person is actually able to be and do to increase her well-being. To capability theorists, 'the freedom to achieve well-being is of primary moral importance'[14], and can therefore be used to evaluate if a person's social, political, and developmental circumstances support or hinder her well-being. In more technical terms, a capability is the real opportunity (or freedom) to achieve functionings, where functionings are beings and doings (or states) of a person, like 'being well-nourished' or 'taking the bus to work'. Both capabilities and functionings are treated as a measure of a person's well-being, and therefore allow us to compare people in terms of how well their life is going. They are distinguished, however, from resources like wealth or commodities, because those metrics arguably provide only limited or indirect information about how well the life of a person is going.

[9] E.g., A Sen, *Commodities and Capabilities* (1985) and A Sen, *Development as Freedom* (2001); as an introduction also cf. A Sen, 'Development as Capability Expansion' (1989) 19 *Journal of Development Planning* 41–58.

[10] E.g., M Nussbaum, *Women and Human Development: The Capabilities Approach* (2001) (hereafter Nussbaum, 'Capabilities Approach'); as an introduction cf. M Nussbaum, *Creating Capabilities: The Human Development Approach* (2013) (hereafter Nussbaum, 'Creating Capabilities').

[11] Nussbaum, 'Creating Capabilities', 18.

[12] C Gore, 'Irreducibly Social Goods and the Informational Bias of Amartya Sen's Capability Approach' (1997) 9(2) *Journal of International Development* 235–250.

[13] SM Okin, 'Poverty, Well-being, and Gender: What Counts, Who's Heard?' (2003) 31(3) *Philosophy & Public Affairs* 280–316.

[14] I Robeyns and M Fibieger Byskov, 'The Capability Approach' (2020) *The Stanford Encyclopedia of Philosophy Winter 2020 Edition* https://plato.stanford.edu/archives/win2020/entries/capability-approach.

Nussbaum further developed the capabilities approach, specifically by extending the scope of *Sen*'s pragmatic and result-oriented theory.[15] For her, a functioning is 'an active realization of one or more capabilities (. . .). Functionings are beings and doings that are the outgrowths or realization of capabilities.'[16] *Nussbaum* stresses that she does not intend to deliver a theory on human nature as such. But she does understand the capabilities approach as an inherently evaluative and ethical theory that focuses on valuable capacities that human beings have reason to value and that a just society is obligated to nurture and support.[17] The normative criterion for valuableness is well-being as well (although quality of life or human flourishing are sometimes used synonymously). According to *Nussbaum*'s ambitious theory, the development of capabilities is connected to the notions of freedom (like in *Sen*'s theory) and dignity (by which she is going beyond *Sen*); she states: 'In general (. . .) the Capabilities Approach, in my version, focuses on the protection of areas of freedom so central that their removal makes a life not worthy for human dignity.'[18] Against this background *Nussbaum* famously compiled a list with ten central capabilities, ranging from life, bodily health, bodily integrity, up to the affiliation with others and the political and material control over one's environment.[19]

Our conception of cyberbilities shares not only *Sen*'s focus on well-being and functionings, but also *Nussbaum*'s idea to provide a list with core cyberbilities. However, we understand our list not as a substitution, but a supplement to *Nussbaum*'s, taking into account that AI-based brain–computer interfaces might change our understanding of both capabilities and agency.

Our reasoning is that modern technology is so complex and closely connected to human agency and well-being that it has the potential not only to subvert, but also to strengthen capabilities in complex ways. This relation will only become more intricate as neurotechnology and AI become more elaborate and integrated in our bodies, especially with human–machine fusions promised by future BCI technologies. Simply asking if such technologies contribute to or detract from well-being, or contradict or strengthen central capabilities, might be undercut by the impact they have on human agency as a whole. We could overlook subtle but unpreferable effects on agency if a technology grants certain well-being benefits, or miss beneficial effects on flourishing, for example in the case of capability-tradeoffs[20] realized by new types of technologically-assisted agency.

For this reason, we argue that evaluating current and future neurotechnology on the basis of the capabilities approach alone might fall short. Instead, we propose to combine the well-being and functioning focus of the capability approach with an extended perspective on agency that is tailored to identifying the impact of neurotechnology and AI on human agency as a whole. The specific challenge is that neurotechnological devices are not just another type of tool that human beings can use as an external means to realize capabilities and achieve well-being. By intervening into the brain of a person, neurotechnology interacts intimately with the basis of human agency, which opens the possibility to affect agency and capabilities in unforeseen ways. Because we may not be able to predict if this new kind of interaction relates positively or negatively to those dimensions, it seems prudent to develop a perspective that may accompany the coming neurotechnological developments with ethical scrutiny.

[15] Nussbaum, 'Creating Capabilities' (n 10).
[16] Ibid, 25.
[17] Ibid, 28.
[18] Ibid, 31.
[19] Ibid, 33–34.
[20] Cf. Section V.

Hence, cyberbilities are an extension of the core tenets of the capability approach insofar as they are capabilities that arise from agency that is already enabled or affected by neuro- and/or AI-technology.

2. *Agency and Human–Machine Interactions*

After having briefly introduced the notion of capabilities we now focus on the conceptions of agency and human–machine interaction. This section will work towards an understanding of the ways in which human agency intersects and merges with machinic and software agency in technological contexts, a phenomenon which sociologist *Rammert* calls distributed agency.[21] The concept of hybrid agency, which we introduce in Section III, is a specific type of distributed agency which is also the core of the notion of a cyberbility.

There are two dimensions we consider to be central to human–machine interaction in general, and human–computer interaction in particular: Firstly, the causal efficacy of intentions, in other words, the idea that human intentions are the causal origin of technologically mediated actions, and secondly, the social aspect of acting in a technological context, especially when interacting with technological devices. We review established views on both agency and human–machine interaction in the context of BCI operation[22] and then go on to discuss these views in more depth.[23] While these two dimensions by no means exhaust the spectrum of relevant aspects in human–machine interaction, we see them as instructive starting points to develop our extended view that leads to introducing the novel concepts of hybrid agency and cyberbilities.

a. The 'Standard View': Compensating Causality and Interactivity

Philosophically speaking, the concept of agency is connected with the phenomenon of intentionality and intention. An intention is a specific type of mental state that aggregates other action-related mental states (such as beliefs and desires), representing a concrete goal or plan and adding a stable commitment to actually perform actions aimed at realizing the respective goal or plan.[24] Theories that explain how intentions work conceptually are numerous[25], but the so-called standard view is that intentions govern and direct behavior through their specific causal efficacy.[26] In other words, intentions govern behavior by virtue of their direct and indirect causal effects on the chain of events from mental states to the execution of movements.[27] Hence, saying that a person 'has agency' amounts to saying that his intentions causally affect how the brain produces behavioral output, from cortical to spinal neural activity.

This view of agency is common not only in philosophy, but also in other disciplines, such as psychology and neuroscience. Principally, these disciplines agree that our behavior is governed by causally efficacious mental states, which emerge from the brain as their physiological basis. As a result, this view is compatible with a neuroscientific view of behavior and agency, and can

[21] Cf. Rammert, 'Distributed Agency' (n 8), 77–86.
[22] Cf. Section II 2(a).
[23] Cf. Section II 2(b) and II 2(c).
[24] Cf. M Bratman, *Intention, Plans, and Practical Reason* (1987).
[25] Cf. T O'Connor and C Sandis (eds), A *Companion to the Philosophy of Action* (2010).
[26] E.g., A Mele, *Springs of Action: Understanding Intentional Behavior* (1992); M Bratman, *Faces of Intention: Selected Essays on Intention and Agency* (1999).
[27] For an account detailing the effects of intentions not only on other mental states but also neurophysiological states underlying the execution of movements, cf. E Pacherie, 'The Phenomenology of Action: A Conceptual Framework' (2008) 107 *Cognition* (hereafter Pacherie, 'Action').

be used to describe the basic rationale of current brain–computer interfaces and neuromodulation technologies. In what follows, we will primarily focus on motoric neuroprostheses, as they provide a clear and instructive case of application. The rationale for motoric neuroprostheses reads: If an agent cannot perform actions and movements anymore because the causal chain from the brain to the extremities is, in some way or another, interrupted, disrupted, or limited the brain–computer interface can bridge causal gaps in this chain by (re-)connecting the neural correlates of intention with an artificial effector, such as a wheelchair or robotic arm.[28]

This basic rationale highlights a mainly restorative and supplemental quality of neurotechnology, which we call the compensatory view, as its main focus is to compensate for lost or limited neural function. The compensatory nature of neurotechnology is illustrated by *Walter Glannon* in his analysis of the specific interaction between brain–computer interface and user. Arguing that neurotechnologically assisted agency is comparable to natural agency, *Glannon* states: 'BCIs do not supplant, but supplement the agent's mental states in a model of shared control. Rather than undermining the subject's control of his behavior, they enable control by restoring the neural functions mediating the relevant mental and physical capacities'.[29] Besides drawing on the standard view of agency, in which the device compensates for the interrupted chain of events from intention to movement by bridging causal gaps, *Glannon* states that an 'extended embodiment'[30] is a further prerequisite: If the user fails to experience the device as part of her own body schema, she may not perceive the movements of the robotic arm as 'her own' which could 'undermine the feeling of being in control of one's behavior', thereby disrupting her sense of agency.[31]

According to *Glannon*, the restorative and supplemental character of brain–computer interfaces stems from the specific interaction between user and device, which creates the phenomenon of shared control, in other words, control over the course of action is partly on the side of the user, and partly delegated to the brain–computer interface. The interaction consists of the user directing her mental states in such a way that the interface can detect neural states which 'encode' her intentions. This kind of interaction is the basis of *Glannon*'s notion of shared control, and successful extended embodiment is necessary to sustain and improve this kind of interactive control.

Brain–computer interfaces based on these principles have been successfully implemented in human patients, and the technology clearly has the potential to compensate for limitations of agency in the way described above. However, it is important to note that while this conclusion is valid, it also stems from a specific understanding of technology, which might support the conclusion while also obscuring other relevant aspects. The compensatory view conceives neurotechnology as a type of instrumental technology and hence frames brain–computer interfaces as auxiliary devices. From this perspective, neurotechnological devices are conceptualized as tools which remain, by definition, fundamentally subordinate to human autonomy and intention. BCI operation appears as auxiliary in nature because the device takes over only partial segments of a course of action, and the overall goal and regulation of action remains governed by human agency.

[28] The same general rationale applies to many other use cases of neurotechnologies that alter, modulate, or monitor brain activity to, for example, enable the use of digital keyboards or cursors, neurofeedback systems, or brain stimulation devices such as deep-brain-stimulators.

[29] W Glannon, 'Neuromodulation, Agency and Autonomy' (2014) 46 *Brain Topography* 27 (hereafter Glannon, 'Neuromodulation').

[30] Ibid, 51.

[31] Ibid, 51. Note that experiencing control over one's behavior may be only one aspect of the sense of agency (cf. Subsection II 2(c)).

In principle, many technological settings can be usefully described from the perspective of instrumental technology. But is this the case in BCI operation? After all, a brain–computer interface is not just an external object, but a device implanted into the brain, affecting and interacting with the origins of action rather than just the external locus of object manipulation. So, does this intimate characteristic distinguish a brain–computer interface from an external tool?

To address this question, we need to examine the effects of BCI operation concerning its causal and neurophysiological nature to see if brain–computer interfaces 'just bridge a causal gap', or if they do more than that.[32] This analysis will suggest that the compensatory view on BCI technology is an extension of the standard view on agency, thereby inheriting its conceptual limits. To counteract this limitation, we need to extend the vocabulary we use to describe agency, and we will do this by taking a closer look at the specific kind of interaction between user and device, taking into account certain social characteristics of this interaction.[33] The basic idea we need to address is that there are some human–machine interactions which are so intimate that it becomes hard to say where human agency ends and machine-agency starts: The interaction between human and machine is such that agency is actually distributed across both interaction partners, rather than ultimately remaining under the governance of human intention.

b. Reframing Causality

Concerning the neurophysiological nature of a brain–computer interface operation, and shared control specifically, it should be noted that 'a brain–computer interface records brain activity' does not mean that it simply 'detects intentions in the brain'. A brain–computer interface is not like an ECG that detects a heartbeat. Rather, operating a brain–computer interface relies on a mutual learning process: Recently developed interfaces increasingly rely on machine learning to distinguish relevant from irrelevant information about intended movement from a narrow recording site that yields a stream of noisy and limited data.[34] At the same time, the user has to learn to influence his neural activity in such a way that the recording site provides enough information in the first place to successfully operate the external effector. This is achieved by passing through a lengthy training period in which user and interface gradually attune and adapt to each other.[35] Shared control over actions in *Glannon*'s sense is based on this kind of mutual adaptation.[36]

However, this attunement and adaptation between brain–computer interface and user also affects the brain as a whole, which mitigates the claim that in these user–computer interactions, control is merely partly delegated from user to device. As *Jonathan R. Wolpaw* and *Elizabeth Winter Wolpaw* note, natural (i.e. not neurotechnologically assisted) agency is a product of activity distributed across the whole central nervous system, which continually adapts and changes to produce appropriate behavioral responses to its environment.[37] Introducing a brain–computer interface basically creates a novel output modality for this complex system.

[32] See Sub-section II 2(b).

[33] See Sub-section II 2(c).

[34] For an overview of the principles of brain–computer interface operation see JR Wolpaw and EW Wolpaw (eds), *Brain-Computer Interfaces: Principles and Practice* (2012) (hereafter Wolpaw and Wolpaw, 'Brain-Computer Interfaces') or B Graimann, B Allison, and G Pfurtscheller (eds), *Brain-Computer-Interfaces: Revolutionizing Human-Computer-Interaction* (2010).

[35] For an exemplary case see JL Collinger and others, 'High-Performance Neuroprosthetic Control by an Individual with Tetraplegia' (2013) 381 *Lancet* 557–564.

[36] Cf. Wolpaw and Wolpaw, 'Brain-Computer Interfaces' (n 34) 7: 'BCI operation depends on the interaction of two adaptive controllers [brain and BCI]'.

[37] Wolpaw and Wolpaw, 'Brain-Computer Interfaces' (n 34) 6.

As a result, the central nervous system as a whole adapts and rearranges in order to learn to control this new way of interacting with its surroundings. And because brain–computer interfaces rely on a localized recording site and a specific type of neural signal, the user needs to retrain a small part of this extensive system to provide an output which normally is produced by the whole central nervous system, which in turn affects how the central nervous system works as a whole.

In our view, this speaks against the basic tenet of the compensatory view that a brain–computer interface just supplements the agent's mental states, as the whole system that is producing mental states is affected by neurotechnological interfacing. Specifically, it puts into question the view that a brain–computer interface simply bridges a causal gap in the action chain of its user, as a brain–computer interface does not carefully target a specific causal gap. Rather, it modulates the whole system to restore causal efficacy, restructuring the causal chain from intention to action in the process. While this does not mean that a brain–computer interface necessarily supplants a person's agency, we still claim that the compensatory view might easily miss important ramifications of the technology, even in terms of causal efficacy. Furthermore, we argue that the compensatory view also falls short of identifying more overarching agency-altering effects of neurotechnology. While *Glannon* discusses aspects of the sense of agency in terms of extended embodiment and experiencing control over one's behavior – important aspects that contribute to explaining the sense of agency – both embodiment and the sense of agency include further aspects. For example, it has been suggested that the sense of agency is an aggregation of at least three distinct phenomena, namely, the sense of intentional causation, the sense of initiation, and the sense of control.[38] The latter can be distinguished further into the sense of motor, situational, and rational control[39], raising the question of which aspects of control are actually shared between user and brain–computer interface. While the case of motor control seems quite clear, any effect of a neurotechnological device on rational or situational control over actions should be analyzed rigorously – the question is if an exclusively causal and neurophysiological vocabulary will suffice to explore these effects and their overarching consequences. It is to be expected that this situation will become even more pressing with the inclusion of increasingly complex and autonomous AI-technology. As outlined earlier, even current machine learning–supported brain–computer interfaces cannot be understood as simple 'translators' between brain and computer. Advanced AI-technologies will likely introduce additional dimensions of influence by establishing more sophisticated means of interaction between human and machine. We argue that this necessitates a framework that can capture not only specific causal effects, but also changes in interactivity between human and machine which might modulate the causal setting of agency altogether.

c. Reframing Interactivity

The compensatory view addresses interactions between user and brain–computer interface by highlighting that both the causal compensation and the integration into the body schema is based on a reciprocal learning process. However, the interactions and adaptations between user and brain–computer interface also have a social dimension which is not addressed by the compensatory view. We argue that this is due to conceptual blind spots that result from its vocabulary, which treats agency and intentionality as purely biological functions. As a result, the compensatory view struggles with identifying and factoring in nonbiological (e.g., social and normative) and nonhuman (i.e. artificially intelligent) dimensions of agency.

[38] Cf. Pacherie, 'Action' (n 27) who integrates empirical studies in her theory. For phenomenological aspects see S Gallagher, 'Multiple Aspects in the Sense of Agency' 31(1) *New Ideas in Psychology*.

[39] Pacherie, 'Action' (n 27) 209–213. Also cf. J Shepherd, 'The Contours of Control' (2014) 170 *Philosophical Studies*.

To counteract this shortcoming, it is necessary to extend the vocabulary of agency accordingly. Sociology, Science, and Technology Studies and Philosophy of Technology have a rich history of analyzing how technology permeates modern life and deeply affects and changes human agency. We will paradigmatically draw on a sociological theory called the gradualized concept of agency[40], which shifts the focus from agency as a biological capacity to agency as a phenomenon that emerges from various types of interactions between and among humans, machines, and software. Advanced technologies, it is argued, create a multitude of heterogeneous artificial 'agencies' which interact and influence not only each other, but also human agency in fundamental ways. Importantly, the gradualized concept of agency can be used to examine interactions between a brain–computer interface and its user on the level of human–machine interactions without contradicting the neurophysiological aspects of human agency discussed earlier. In fact, the gradualized concept of agency may help to emphasize that the compensatory view is not outright false by demonstrating its blind spots in a constructive manner.

As argued above, the compensatory view regards neurotechnology as a passive tool by arguing that its contributions to a course of instrumental action concern only partial sequences in the causal chain, while the order of causal events still is governed and regulated by human intention. Hence, the significance and involvement of technological contributions is derived primarily from human intention: The user and his intentions remain in control of the action.

By contrast, the gradualized concept of agency offers an analysis of this kind of relation that shows how advanced technology can subtly restructure instrumental action and lead to agency-altering consequences. It draws on an action-theoretic distinction between three dimensions of agency. The intentional dimension contains the rational capacity to set action goals and deliberate courses of action. Human intention embodies this capacity as an overarching mental state that governs action from planning to execution. The regulative dimension corresponds to control and monitoring of action courses. And the effective dimension describes the base level efficacy to causally affect the environment depending on intentional and regulative aspects.[41]

Based on this model, the gradualized concept of agency argues that technological involvement in the effective dimension can easily cascade from the effective to the regulative and even the intentional dimension. Three common motives of instrumental action illustrate this shift, as technology is often used to delegate effective and regulative aspects of actions in order to save time, improve action outcomes, and to realize action goals the agent could not realize herself. While these aspects may not seem noteworthy when using a conventional tool like a hammer or a common car, their significance and interconnectivity increases the more advanced a technological device is. This can be illustrated by way of two examples: Firstly, a navigation system not only saves time when planning a route, it also improves travel times by calculating and continuously adjusting the best route based on actual traffic data; and secondly, the Google search algorithm seems to be a simple tool to search for relevant information on the Internet. But by scanning billions of websites and documents in fractions of a seconds it is not only infinitely more efficient in finding information, but also autonomously regulates the search by ranking

[40] Cf. W Rammert and I Schulz-Schaeffer, 'Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Abläufe verteilt' in W Rammert and I Schulz-Schaeffer (eds) *Können Maschinen handeln?* 11–64 and I Schulz-Schaeffer and W Rammert. 'Technik, Handeln und Praxis. Das Konzept gradualisierten Handelns revisited' in C Schuber and I Schulz-Schaeffer (eds) *Berliner Schlüssel zur Techniksoziologie* 41–76. For further aspects also see I Schulz-Schaeffer, 'Technik und Handeln. Eine handlungstheoretische Analyse' in C Schuber and I Schulz-Schaeffer (eds) *Berliner Schlüssel zur Techniksoziologie* (hereafter Schulz-Schaffer, 'Technik und Handeln') and Rammert, 'Distributed Agency' (n 8).

[41] Schulz-Schaffer, 'Technik und Handeln' (n 40) 4–5. For an English version with slightly different terminology and line of argument cf. Rammert, 'Distributed Agency' (n 8) 74–77.

relevant information depending on context, which it determines dynamically. Google not only finds information; it evaluates which information is relevant.

It is noteworthy that technological artifacts themselves are the product of complex intentional actions, and that they embody the intentionality of their design: They are 'objectively materialized structures of meaning'[42]. In this perspective, artifacts carry normative weight which affects the structure of the actions they are involved in. Their designed versatility stems from being oriented towards typical rather than individual action, making them multipurpose and offering reliable repeatability of action. As a consequence, using an artifact requires that the agent adapts to its purpose rather than the other way around – particularly in cases where the artifact takes on partial actions which a human agent could not perform. This characteristic illustrates that technology not only improves or creates new courses of action, but that it is suggestive of certain action goals. Hence, artifacts have an active role in the intentional dimension as well. This effect is magnified when artifacts use software algorithms so that the user can delegate aspects of planning, monitoring, and control to the respective program.

These examples show that many interactions between user and advanced technology consist in various forms of delegation. In the context of AI-based neurotechnology, the combination of machines and software is of critical importance, as the involvement of machine learning and other AI-technology amounts to the inclusion of increasingly autonomous software agents in the equation which are capable of the self-generation of actions. Because software agents not only interact with human users, but also (and mostly) with other software agents, their 'intra-activities'[43] create open systems which lose the transparency of operation we usually expect from technological tools. Hence, when delegating actions to such intra-acting software agents, we do not use a tool, but interact with another type of agency. *Rammert* notes that '[w]hen human actions, machine operations and programmed activities are so closely knit together that they form a "seamless web", [we need to] analyze this hybrid constellation as a heterogeneous network of activities and interactivities.'[44] The gradualized concept of agency enables this kind of analysis by proposing the concept of distributed agency, which can be seen as a nondualist perspective[45] on the complex interactions between human and nonhuman contributors to agency. Of particular interest to us is the notion that agency can be (and often is) distributed across a hybrid constellation of entities, including (but not limited to) humans, machines, software, and AI. In this respect, being 'distributed' means that a simple observable movement performed by a patient with a BCI-enabled prosthesis is the result of a complex interplay of activities, interactivities, and intra-activities. So, who is acting in scenarios of neurotechnologically assisted agency? Following the gradualized concept of agency, not a singular agent, but a hybrid constellation of people, machines, and programs over all of which agency is distributed in complex ways.

The concept of distributed agency includes a further dimension which is of importance to our argument, namely the modern sociotechnological setting, or the 'technological condition' we mentioned in the introduction. With the concept of distributed agency, the gradualized concept

[42] Schulz-Schaeffer, 'Technik und Handeln' (n 41) 8, 18–19.

[43] In the gradualized concept of agency, intra-activity describes interactions among artificial (e.g., machinic and software) agents.

[44] Rammert, 'Distributed Agency' (n 8) 82. Note that the gradualized concept of agency defines interactivity as the specific case when human and nonhuman agencies intersect (ibid, 71).

[45] The traditional dualist or asymmetrical perspective on human–machine interaction asserts a dichotomy between 'human action' and 'machine operation', matching the former with the realm of autonomy and morality and the latter with heteronomy and causality (cf. instrumental theories of technology and the paradigm of tool use). The gradualized concept of agency directly opposes this perspective, at least in the case of complex technology.

of agency argues that technologically assisted agency emerges from 'many loci of agency'[46] rather than from singular instrumental actions (e.g., tool use) performed by an individual human agent. While the individual agent does contribute to agency, his contribution is only one activity in a stream of human interactions, machinic intra-activities, and human–machine interactivities. The sociotechnological setting can be addressed by further analyzing human interactions and machinic intra-activities.

Rammert notes that complex technological actions, such as flying tourists to Tenerife with a commercial airplane, include not only individual actions by the pilot, but also considerable contributions from a multitude of both human and nonhuman contributors.[47] On the human side, the pilot is fully dependent on the flight team on board (co-pilot) and on the ground (air traffic controllers, radio operators), as well as the airline company which planned and scheduled the flight, and also the passengers buying the tickets, and so on. On the technical side, the flight is also facilitated by the intra-activities of the various machines and programs integrated into the airplane as well as the respective facilities on the ground. Also, consider that the majority of the flight actions are performed by the auto-pilot, which consists of software programs which constantly measure, monitor, and adjust the mechanical parts of the airplane while checking back with the software networks on the ground which assist in planning, controlling, and navigating the airplane.

Coming back to the example of a movement performed by a patient with an AI-based, BCI-enabled prosthesis, we can apply the same perspective. At first glance, it is just the patient who directly performs the movement of the prosthesis. However, we need to acknowledge the different teams involved, for example, doctors and nurses who performed the initial surgery, and the researchers, technicians, and engineers who built the prosthesis, designed the clinical study, and maintain the device. Also, the hospital, healthcare system, and research and development are related associations of people. And lastly, funding agencies, policies, and social demands contribute to enabling the movement of the neuroprosthesis as well. On the technical side, a neuroprosthesis includes the 'decoder' which can be considered a piece of AI as it employs machine learning to interpret the neural data monitored by the implanted electrodes. While a science fiction example at the moment, the inclusion of more complex AI solutions in brain–computer interfaces may well be achievable in the near future.

III. HYBRID AGENCY AS THE FOUNDATION OF CYBERBILITIES

The concept of distributed agency is a valuable tool to describe agency beyond the scope of the individual biological functions which underlie the human capacity to act in accordance with their intentions and plans. It shifts the perspective from the limited compensatory view of technological agency to the complex context in which technological agency not only takes place but emerges as the product of a broad spectrum of biological, psychological, social, and political factors. In this sense, the notion of distributed agency can be used as a viable philosophical tool to expose the conditions of possibility regarding concepts such as intention or capability.

1. *Distributed Agency and Hybrid Agency*

Because we aim to focus this critical potential on neurotechnologically-assisted agency in particular, we are faced with the challenge to address both its neurophysiological dimension –

[46] Rammert, 'Distributed Agency' (n 8) 78–81.
[47] Ibid, 78–80.

because neurotechnological devices are directly 'wired' into a person's brain – and the socio-technological dimension – as such a device entails complex inter- and intra-activities between and among humans and machines. Thus, we introduce the concept of hybrid agency as a special case of distributed agency, namely as human–machine interactions in which agency is distributed across human and neurotechnological elements. This further emphasizes that neurotechnology – which, by definition, is technology that is directly connected to the brain – is not a conventional tool because it shapes agency not only by being used, but also by directly interacting with the origin of agency. Hence, hybrid agency describes intimate 'fusions' of human and machinic agency and requires direct human–neurotechnology interaction as a basis – but, of course, this does not exclude any biological, psychological, social, or political factors which are directly or indirectly related to neurotechnology as well. These related or indirect factors still shape the structures of neurotechnologically assisted agency, and can themselves be shaped by neurotechnology. And, importantly, hybrid agency specifically includes the various systems of intra-activities among technological and software-agents which neuro-technological devices imply.

The concept of hybrid agency directly opposes the compensatory view, which reduces these complex dimensions by drawing on the instrumental theory of technology, equating neuroprosthetics with conventional tool-use. In this model, neurotechnologically-assisted agency means that a single human agent uses a passive technological tool that compensates for limitations in the action chain, allowing the user to perform actions she would have performed anyway if she could have done so.

2. *Cyberbilities As Neurotechnological Capabilities*

Hybrid agency is the foundation of cyberbilities insofar as this kind of technologically-assisted agency creates specific types of capabilities (i.e. opportunities to gain functionings) which we call cyberbilities. A formal definition reads: 'cyberbilities are capabilities that originate from hybrid agency, i.e. human–machine interactions in which agency is distributed across human and neurotechnological elements.' Because capabilities are defined as real opportunities to achieve functionings – beings and doings that increase well-being – cyberbilities are real opportunities to achieve such functionings as the result of hybrid agency.

It is important to emphasize that cyberbilities are capabilities, not functionings. They are not specific skills or abilities a person may gain from neurotechnology. Rather, they denote the opportunities to gain all kinds of (neurotechnological or 'natural') functionings. And even functionings are not just skills or abilities (doings), but also include states of being (like having financial or social resources or being informed about a certain subject matter). If a paraplegic person uses a brain–computer interface to gain the ability to control her wheelchair, the resulting cyberbilities are related to the opportunities that are gained by this type of technological agency. The brain–computer interface opens up a spectrum of agency that was previously restricted, allowing this person, for example, to attend a wedding and thus participate in socializing, which potentially increases this person's well-being.

Hence, cyberbilities denote the opportunities opening up for users of neurotechnology. But because they are the result of hybrid agency, they are also the product of a technology that affects agency as a whole, in other words, not only on the level of causal efficacy, but also concerning psychological, social, and political factors. While a neurotechnological device may be designed to restore, facilitate, or enhance specific skills, gaining or regaining such skills has wider implications in that this can change how we conceptualize and live our lives. This is why

neurotechnological agency cannot be reduced to gaining specific skills. We devised cyberbilities as a conceptual tool to reflect this important factor and provide a means of orientation concerning the potential developments entailed by the use of neurotechnology. Furthermore, cyberbilities are also concerned with the social ramifications of neurotechnological agency. The more the availability of neurotechnology increases, the more it affects all members of society.

IV. CYBERBILITIES AND THE RESPONSIBLE DEVELOPMENT OF NEUROTECHNOLOGY

After having developed the concept of cyberbilities, we would like to propose a first tentative and incomplete list of cyberbilities, inspired by *Nussbaum*'s list of *capabilities*.[48] We consider our list to be incomplete because it is not meant to cover all basic needs of human beings, nor does it include any other holistic ambition. Therefore, the list presented in the following section should not be understood as a replacement of *Nussbaum*'s list. Rather, we merely aim to stimulate discussions about the implications of future neurotechnologies by drawing on core ideas of the capabilities approach. However, cyberbilities are comparable to capabilities in the following way: *Nussbaum*'s central capabilities describe opportunities which are based on personal and social circumstances which, if restricted or unattainable, would greatly reduce a person's chances to gain well-being-related functionings (to 'lead a good life'). Similarly, cyberbilities describe opportunities created by hybrid agency, which, if restricted or unattainable when using neurotechnology, would greatly reduce the chances to gain well-being-related functionings for a neurotechnologically assisted agent.

Our list of cyberbilities is also necessarily tentative: In order to address future neurotechnologies we have to work with a hypothetical view of neurotechnology that includes a type of AI-supported human–machine fusion that is yet to come. We base this view on current developments, where we can observe various endeavors aiming at advancing AI-assisted neurotechnology, from neuroprostheses for severely paralyzed patients, to sophisticated machine learning approaches, up to straightforward futuristic visions such as *Musk*'s neurotech company Neuralink.[49] Based on such enterprises we think of a future technology that is highly invasive and uses AI methods to generate a novel kind of human–machine fusion that goes far beyond traditional technological tools or machines. We assembled this list with this kind of future technology in mind. In the following, we first introduce our list of cyberbilities,[50] then provide some remarks on the responsible development of neurotechnology,[51] and finally discuss a potential objection against our proposal.[52]

1. *Introducing a List of Cyberbilities*

The five cyberbilities we introduce below fall on a spectrum that ranges from individual to social and political agency. While neurotechnological interventions can create specific neurotechnologically

[48] Cf. Nussbaum, 'Capabilities Approach' (n 10) 78–80.

[49] As a first application, Neuralink wants to develop brain–computer interfaces for patients with spinal cord injury, allowing them to control computers and mobile devices. Neuralink's vision includes constructing an automated robotic neurosurgery system that implants a fully integrated brian–computer interface with over 1000 channels for monitoring and stimulating neuronal activity in multiple brain regions. Neuralink ultimately wants to make this technology available for commercial use (cf. https://neuralink.com).

[50] See Sub-section IV 1.

[51] See Sub-section IV 2.

[52] See Section V.

enabled functionings, they also affect a person in more general ways. New, enhanced, or restored functionings extend and shift a person's individual range of agency, and invasive or otherwise intimate interactions between human and machine may change how a person relates to their body. Both aspects can affect the identity and self-expression of a person, modulating their individual agency. But hybrid agency also affects social agency: On the one hand, neurotechnologies enable individual actions which can be the basis of social interactions and participation, potentially adding a social dimension even to the most basic movements.[53] On the other hand, hybrid agency itself is a type of interaction between human and neurotechnology which already includes various social aspects. Neurotechnology has the potential to support social agency, but some of its aspects may also radically reshape social engagement. Furthermore, hybrid agency has distinct political dimensions that range from enabling a person to take part in communal to political and democratic processes.

Autonomy and self-endorsement: Neurotechnological devices are often used with the intent to restore or increase a person's functionings (skills, abilities, states), which might also suggest that such devices generally support their autonomy as a more general capability. However, this view might be too simplistic if those functionings result from hybrid agency. Hybrid agency entails a relational dimension of autonomy because autonomy is no longer restricted to interactions between human beings but also concerns the interactivity between human and machine. A neurotechnologically-assisted person could retain autonomy in relation to human interactions while losing it in the context of human–machine interaction. Furthermore, due to the intimate fusion of human and machine, simply insisting that the human part must retain autonomy over the machinic part might be an oversimplified demand. Instead, we should address autonomy in this setting not in terms of the primacy and efficacy of human intention (i.e. the compensatory view), but in terms of 'self-endorsed agency'. Autonomy then denotes the extent to which a person experiences their behavior as volitional and self-endorsed as opposed to coerced, driven, or covertly directed by external forces. Understanding autonomy as a cyberbility that is focused on self-endorsed agency might be a viable way to safeguard and promote self-expression and identity.

Embodiment and identity: A technological device should restore or enhance a person's body in such a way that the person is able to integrate the device into her bodily experience, meaning that the person can, without disruptions, identify with the artificial 'part' of herself. She should be able to say 'I have acted like this with the support of the technology' or 'the device and I have acted together' or 'I have acted like this, and I did not experience the interference of the device', etc. Although a neurotechnological device may not be unperceivably 'merged' with the body (like, for instance, a deep brain stimulator), but rather remains separate from the body, the person should have the impression that the device 'behaves' in such a way that she can unreservedly identify with the actions she is performing with the support of the respective device. In other words: The person may not have a sense of ownership but should have a sense of agency. The technological tool should be integrated in the *body schema* of a person, even if the body image is radically changed, for example, in the case of neuroprostheses consisting of external artificial limbs which are 'wired' directly into the motor cortex while remaining clearly separated from the patient's body.

Understandability and life-world: Hybrid agency describes the fusion between a person and a neurotechnological device that is intimately connected with the brain and body of its user. Although a lay person may never entirely comprehend how such a device works exactly, a certain degree of understanding is indispensable. Complementing existing approaches to an

[53] Cf. W Wang and others, 'An Electrocorticographic Brain Interface in an Individual with Tetraplegia' (2013) 8(2) *PLoS ONE*; supplemental material shows the patient controlling an external robotic arm with a brain–computer interface and intentionally touching the hand of his girlfriend for the first time in years: UPMC, 'Paralyzed Man Moves Robotic Arm with His Thoughts' (*YouTube*, 7 October 2011) www.youtube.com/watch?v=yff20TlHv34&ab_channel=UPMC.

'explainable AI', a technological device should be 'understandable' in the sense that the user knows *that* the device creates a situation of hybrid agency and roughly *how* the device might affect her agency and behavior (e.g., knowing that a brain–computer interface complements the causal efficacy of her intentions and where the causal contribution lies, which might concern not only the execution of movements but also their planning or initiation). Furthermore, a person should be able to act in interplay with the device in such a way that she can always identify herself with the resulting joint action. While she does not need to be able to explain how the device works on a technical level, she rather needs to understand how the device contributes to hybrid actions and how the device creates well-being opportunities and, thus, becomes deeply integrated in the person's 'life world'.

Social embeddedness and social experience: Hybrid agency can create opportunities to engage with the social world, be it on the level of restoring mobility and allowing a person to meet other people or on the level of being able to express thoughts and feelings, for example via digital communication devices. Enabling, restoring, and extending such engagements – for example, in the case of severe paralysis, situations that restrict direct social contact (such as a pandemic), or when trying to socialize over long distances – hold the potential of significant well-being gains. At the same time, however, neurotechnology shapes and alters the basic conditions of social interactions, thereby influencing the way both neurotechnology users and nonusers are socially embedded in the first place. One possible way to capture such fundamental changes could be to focus on how our social experiences are affected by technology.

Political engagement and participation: By supporting individual and social agency, neurotechnology also opens up opportunities to engage in political activities on various levels and other forms of campaigning for the common good. Neurotechnological devices should be designed to foster participation in democratic processes such as voting, politicking, or running for office, and should also support engagement in local and global communities, organizations, and institutions.

2. *Remarks on the Responsible Development of Neurotechnology*

Because neurotechnologies are developed within a society and its always changing and shifting norms and regulations, cyberbilities are also linked to broad and ongoing societal, ethical, and legal questions. The keywords listed below are not to be understood as cyberbilities, but as indicators of more general questions surrounding cyberbilities. For example, due to usually limited resources we may encounter questions like which patient would benefit from this technology, meaning that not all persons may have the chance to alter their agency by gaining cyberbilities. Also, the neurotechnological engagement in certain activities may require laws that protect the user's personal data (e.g., online services, healthcare, marketing). Because neurotechnology is and will most likely continue to be heavily regulated, the use of neurotechnology on the individual and social level will inherit the legal and political aspects associated with the regulation of neurotechnology, potentially affecting neurotechnology users and their agency. These complex areas will require careful analysis in the coming years and the following remarks address some of the most basic requirements to safeguard the responsible development of neurotechnology. Furthermore, both the question of the trustworthiness of technological devices (especially regarding AI systems) in general and questions around data protection and informational self-determination will affect the future of neurotechnology and also how we evaluate cyberbilities in the future.

Availability: Market approval of neurotechnological devices is related to a host of important questions. Who will have access to neurotechnology? How is access regulated – via healthcare systems, or even the open market? And how does regulated access affect not only neurotechnology

users, but also those who do not have access to neurotechnology and who have to interact or compete (e.g. in the job market) with those who do? Such questions indicate important consequences for well-being on multiple levels: If neurotechnology users are individually, socially, politically, or otherwise advantaged or disadvantaged, this circumstance generally affects neurotechnology-related opportunities to gain well-being – both for those who have and those who do not have access to neurotechnology. The question of availability specifically reveals that neurotechnology affects not only those who gain hybrid agency, but also those who do not. This aspect could even result in a 'feedback loop', as the relationship between neurotechnology users and nonusers might affect how norms and regulations develop, further changing this initial relation.

Data protection: Because neurotechnological devices monitor, record, and process neurophysiological (and potentially other biological or psychological) data, hybrid agency opens up a plethora of ways in which the data can be used and shared to create functionings or cyberbilities. But the same data could also be used for, among other things, political or commercial purposes. A neurotechnological device should be designed in such a way that it collects and uses personal data as conservatively as possible (e.g. restricted to momentary joint actions and activities), or at least implements particularly robust measures to prevent misuse of data (e.g. through encryption). Because AI (i.e. machine learning) is already implemented in neuroprostheses in order to interpret brain activity faster and more efficiently, such devices should be regarded as a genuine 'part' of the patient and thus be subject to the same legal and political protection concerning personal information and human rights as the user herself. Also, any further implementation of AI-technology needs to be carefully designed to safeguard both the data of its user and any human or nonhuman interaction partners.

Trustworthiness: A technological device should not only be reliable in a mere technological sense, but the person should be able to trust herself and the device, especially in cases when the device is merged with the human body or brain. This trust could be seen as a broad psychological foundation of neurotechnology usage, as it includes many of the other items on this list and the list of cyberbilities, like trusting that hybrid agency can be self-endorsed, confidence in the physiological safety and digital security (hacking, manipulation, privacy) of neurotechnology, and reliance on understanding, in principle, the ways in which the device modifies and influences one's natural capacity for agency.

V. DISCUSSION AND CLOSING REMARKS

Neurotechnology will continue to afford us with astounding possibilities. While the application of neurotechnology is currently restricted to medical usage, we hope that we provided a convincing argument anticipating the future scope of this technology going above and beyond the therapeutic restoration of specific skills and abilities. The proposition of the concepts of hybrid agency and cyberbilities is directed at broadening our perspective so that the enormous potential and overarching impact of neurotechnology may come to the fore.

However, we want to discuss one objection that could be raised on this point, namely that the focus on well-being is too one-sided and may lead to disregarding the intrinsic value of human agency. After all, cyberbilities are not based on 'natural' agency, but hybrid agency. What if this novel kind of agency is in some way deficient, because its technological portion somehow detracts from the human part of agency? In some cases, then, well-being could be achieved at the price of losing aspects of 'natural' agency.

This reasonable objection raises questions about the relative normative weights of well-being and agency, a topic that also applies to the capability approach. There, capabilities and functionings are embedded in the more general concept of agency, and the latter itself has an intrinsic normative

value. But does the importance of agency outweigh the importance of well-being? If we transfer this question to the cyberbilities approach, we could ask: Could the pursuit of cyberbilities lead to justifying a loss of 'natural' agency for the sake of gaining well-being that is less connected to human agency, but rather grounded in technological agency? And to add a utopian twist, could an AI-based brain – computer interface at some point know better and decide itself whether human or technological agency leads to more well-being gains?

There probably is no clear answer to these questions. While it could be argued that this thought experiment warrants preserving 'natural' agency, our line of argument in previous sections hopefully demonstrated that 'natural agency' is not easy to define. Following the standard view, natural agency would mean that the intentions of the human agent systematically modulate which actions are carried out. But considering the gradualized concept of agency, we also saw that human agency is entangled in complex social, institutional, and political systems that influence which intentions are available to human agents in the first place. Human agency is already intrinsically affected by our use of technology and its sociopolitical context.

However, we want to address a point we think is related to this general question: the possibility of capability-tradeoffs. We argued that neurotechnology might not just compensate for causal gaps in the action chain, but rather has an influence on the entire action chain by modulating how the brain works as a whole. Furthermore, neuroechnology also affects, in various ways, the formation of intentions that lead to action chains in the first place. As a result, neurotechnology has the potential to lead to both gaining and losing capabilities.

Consider this example: A neurotechnological device might allow a person to achieve mobility-based functionings (like performing grasping movements with a robotic arm, or getting to work with a wheelchair controlled with the help of a brain–computer interface). If this device also has the effect that its user does not experience her movements as caused by herself (significant portions of grasping movements are controlled by the prosthesis; the wheelchair autonomously navigates to the workplace), then the well-being achievements (being self-sufficient at home and earning money) are realized at the cost of losing some portion of agency. This is a capability tradeoff: The capability (in this case, cyberbility) of neurotechnologically enabled mobility is traded off against the capability of controlling and planning one's movements (which is a part of 'natural' agency).

Of course, such tradeoffs are not necessarily adverse or harmful: In the case of grasping, delegating control to the device at the cost of the sense of control might be acceptable as long as a general sense of agency remains intact (for instance, if the prosthesis overall performs in line with the user's intentions). The case of the autonomous wheelchair is similar, although here the delegation of control goes much further because it includes planning and deciding how to navigate. Our argument is, there might be a point at which the 'cost' becomes unacceptable, for example, if significant portions of agency are traded off. Possible examples could be that the device increasingly detracts from agency, severely influences the decisions of users, or significantly affects the process of intention formation.

Naturally, determining the point at which capability tradeoffs become unacceptable is a difficult task as this is not a technical or scientific problem, but a normative one that needs to be addressed from ethical, legal, social, and political viewpoints. But this open question might help to conclude our line of argument, as we understand cyberbilities as a potential safeguard against unacceptable capability tradeoffs.[54]

[54] Funding acknowledgement: The work leading to this publication was supported by FUTUREBODY, funded by ERA-NET NEURON JTC2017.

PART VIII

Responsible AI for Security Applications and in Armed Conflict

26

Artificial Intelligence, Law, and National Security

Ebrahim Afsah

I. INTRODUCTION: KNOWLEDGE IS POWER

The conjecture 'that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it'[1] has motivated scientists for more than half a century, but only recently attracted serious attention from political decision-makers and the general public. This relative lack of attention is perhaps due to the long gestation of the technology necessary for that initial conjecture to become a practical reality. For decades merely an aspiration among a small, highly skilled circle engaged in basic research, the past few years have witnessed the emergence of a dynamic, economically and intellectually vibrant field.

From the beginning, national security needs drove the development of Artificial Intelligence (AI). These security needs were motivated in part by surveillance needs, especially code-breaking, and in part by weapons development, in particular nuclear test simulation. While the utilisation of some machine intelligence has been part of national security for decades, the recent explosive growth in machine capability is likely to transform national and international security, consequently raising important regulatory questions.

Fueled by the confluence of at least five factors – the increase in computational capacity; availability of data and big data; revolution in algorithm and software development; explosion in our knowledge of the human brain; and existence of an affluent and risk-affine technology industry – the initial conjecture is no longer aspirational but has become a reality.[2] The resulting capabilities cannot be ignored by states in a competitive, anarchic international system.[3]

[1] As succinctly put in the project proposal to the 1956 Dartmouth Conference; J McCarthy and others, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955' (2006) 47 *AI Magazine* 12.

[2] NJ Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (2010) (hereafter Nilsson, The Quest for Artificial Intelligence).

[3] The literature is extremely copious, a good point of departure is H Bull, *The Anarchical Society: A Study of Order in World Politics* (1977); KA Oye, 'Explaining Cooperation under Anarchy: Hypotheses and Strategies' (1985) 38 *World Politics* 226. Professor Oye was the convener of the talk by Judge James Baker at MIT on 6 March 2018 that initially got me interested in AI, my intellectual debt to his work is gratefully acknowledged. See JE Baker, 'Artificial Intelligence and National Security Law: A Dangerous Nonchalance' (2018) 18-01 MIT Starr Forum Report (hereafter Baker, 'Artificial Intelligence and National Security Law').

As AI becomes a practical reality, it affects national defensive and offensive capabilities,[4] as well as general technological and economic competitiveness.[5]

There is a tendency to describe intelligence in an anthropomorphic fashion that conflates it with emotion, will, conscience, and other human qualities. While this makes for good television, especially in the field of national security,[6] this seems to be a poor analytical or regulatory guideline.[7] For these purposes, a less anthropocentric definition is preferable, as suggested for instance by *Nils Nilsson*:

> For me, artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment. According to that definition, lots of things – humans, animals, and some machines – are intelligent. Machines, such as 'smart cameras,' and many animals are at the primitive end of the extended continuum along which entities with various degrees of intelligence are arrayed. At the other end are humans, who are able to reason, achieve goals, understand and generate language, perceive and respond to sensory inputs, prove mathematical theorems, play challenging games, synthesize and summarize information, create art and music, and even write histories. Because 'functioning appropriately and with foresight' requires so many different capabilities, depending on the environment, we actually have several continua of intelligences with no particularly sharp discontinuities in any of them. For these reasons, I take a rather generous view of what constitutes AI.[8]

[4] M Karlin, 'The Implications of Artificial Intelligence for National Security Strategy' in A Blueprint for the Future of AI (Brookings, 1 November 2018) www.brookings.edu/series/a-blueprint-for-the-future-of-ai/; A Polyakova, 'Weapons of the Weak! Russia and AI-Driven Asymmetric Warfare' in A Blueprint for the Future of AI (Brookings, 15 November 2018) www.brookings.edu/series/a-blueprint-for-the-future-of-ai/; M O'Hanlon, 'The Role of AI in Future Warfare' in A Blueprint for the Future of AI (*Brookings*, 29 November 2018) www.brookings.edu/series/a-blueprint-for-the-future-of-ai/.

[5] Much current attention is given to China's single-minded pursuit of attaining technological competitiveness by 2025 and leadership by 2035, including in the field of AI. The State Council published in July 2017 a 'New Generation Artificial Intelligence Development Plan' that built on the May 2015 'Made in China 2025' plan, which had already listed 'new information technology' as the first of ten strategic fields. The two plans are accessible at https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/ and http://english.www.gov.cn/2016special/madeinchina2025/. For a discussion see *inter alia* 'AI in China' (*OECD*, 21 February 2020) https://oecd.ai/dashboards/countries/China; 'AI Policy China' (Future of Life Institute, February 2020) <https://futureoflife.org/ai-policy-china/; P Mozur and SL Myers, 'Xi's Gambit: China Plans for a World without American Technology' *New York Times* (11 March 2021) www.nytimes.com/2021/03/10/business/china-us-tech-rivalry.html (hereafter Mozur and Myers, 'Xi's Gambit'); X Yu and J Meng, 'China Aims to Outspend the World in Artificial Intelligence, and XI Jinping Just Green Lit the Plan' *South China Morning Post* (18 October 2017) www.scmp.com/business/china-business/article/2115935/chinas-xi-jinping-highlights-ai-big-data-and-shared-economy.

[6] Perhaps most enduringly in the 1983 movie 'WarGames', where a recently commissioned intelligent central computer is hacked into by a teenager, who inadvertently almost causes nuclear Armageddon. This is only averted when the computer learns, after playing Tic-Tac-Toe with the teenager, that nuclear war cannot have a winner, causing him to rescind the launch command and to comment: 'A strange game. The only winning move is not to play.' There are obvious allusions to the doomsday machine scenario discussed further below. Interestingly, simultaneous to the film but unbeknownst to most until much later, the automated early warning system of the Soviet Union on 26 September 1983, at a time of extreme tension between the two countries, falsely indicated an American nuclear attack, almost triggering a catastrophic retaliatory nuclear attack. This was stopped by Lieutenant Colonel Stanislav Petrov, who disobeyed orders because he intuited that it was a false alarm; M Tegmark, 'A Posthumous Honor for the Man Who Saved the World' (*Bulletin of the Atomic Scientist*, 26 September 2018) https://thebulletin.org/2018/09/a-posthumous-honor-for-the-man-who-saved-the-world/.

[7] A Chayes, 'Cyber Attacks and Cyber Warfare: Framing the Issues' in A Chayes (ed), *Borderless Wars: Civil Military Disorder and Legal Uncertainty* (2015) (hereafter Chayes, 'Cyber Attacks and Cyber Warfare'); L DeNardis, 'The Emerging Field of Internet Governance' (2010) Yale Information Society Project Working Paper Series https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1678343 (hereafter DeNardis, 'The Emerging Field of Internet Governance').

[8] Nilsson, *The Quest for Artificial Intelligence* (n 2) xiii.

The influential *Stanford 100 Year Study on Artificial Intelligence* explicitly endorses this broad approach, stressing that human intelligence has been but the inspiration for an endeavour that is unlikely to actually replicate the brain. It appears that intelligence – whether human, animal, or machine[9] – is not necessarily one of clearly differentiated *kind*, but ultimately a question of *degree* of speed, capability, and adaptability:

> Artificial Intelligence (AI) is a science and a set of computational technologies that are inspired by – but typically operate quite differently from – the ways people use their nervous systems and bodies to sense, learn, reason, and take action. . . . According to this view, the difference between an arithmetic calculator and a human brain is not one of kind, but of scale, speed, degree of autonomy, and generality. The same factors can be used to evaluate every other instance of intelligence – speech recognition software, animal brains, cruise-control systems in cars, Go-playing programs, thermostats – and to place them at some appropriate location in the spectrum.[10]

At its most basic, AI means making sense of data, and can thus be differentiated from cyberspace, which primarily concerns the transmission of data. Collecting data is fairly inconsequential without someone to analyse and make sense of it.[11] If the purpose of a thought or action can be expressed numerically, it can be turned into coded instructions and thereby cause a machine to achieve that purpose. In order to understand the relationship better, it is helpful to differentiate between data, information, knowledge, and intelligence.

Data is raw, unorganised, factual, sensory observation, collected in either analog or digital form, with single data points unrelated to each other. Already in this raw form, data can be used by simple machines to achieve a purpose, for instance temperature or water pressure readings by a thermostat switching a heater on or off, or a torpedo's depth sensor guiding its steering system. Observed and recorded facts can take many forms, such as statistics, satellite surveillance photographs, dialed phone numbers, etc. Such data, whether qualitative or quantitative, stands on its own and is not related to external signifiers. In this form, it is not very informative and fairly meaningless. Where analog storage is logistically limited, the recording of observational data in electronic, machine-readable form is no longer physically limited.

Information, by contrast, depends on an external mental model through which data acquires meaning, context, and significance. Data becomes information through analysis and categorisation; it acquires significance only through the imposition of order and structure. Information is, therefore, data that has been processed, organised according to meaningful criteria, given context, and thereby made useful towards achieving outcomes according to predetermined

[9] Human denial of both intelligence and consciousness in other creatures seems ultimately to be a fairly straightforward case of cognitive dissonance: 'To me, consciousness is the thing that feels like something,' said Carl Safina, an ecologist. 'We're learning that a lot of animals – dogs, elephants, other primates – have it. . . . I think it's because it's easier to hurt them if you think of them as dumb brutes. Not long ago, I was on a boat with some nice people who spear swordfish for a living. They sneak up to swordfish sleeping near the surface of the water and harpoon them, and then the fish just go crazy and kind of explode. When I asked, 'Do the fish feel pain?' the answer was, 'They don't feel anything.' Now, it's been proven experimentally that fish feel pain. I think they feel, at least panic. They clearly are not having a good time when they are hooked. But if you think of yourself as a good person, you don't want to believe you're causing suffering. It's easier to believe that there's no pain.' C Dreifus, 'Carl Safina Is Certain Your Dog Loves You' *New York Times* (21 October 2019) www.nytimes.com/2019/10/21/science/carl-safina-animal-cognition.html.

[10] 'Artificial Intelligence and Life in 2030 – One Hundred Year Study on Artificial Intelligence, Report of the 2015 Study Panel' (*Stanford University*, September 2016) 4, 12 https://ai100.stanford.edu/2016-report.

[11] T Zarsky, '"Mine Your Own Business!": Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion' (2003) 5 *Yale J L & Tech* 1, 4 *et seq* (hereafter Zarsky, 'Mine Your Own Business!').

needs. This process is dependent on the existence of conceptional models created in response to these needs.[12] Significance, meaning, and usefulness are, therefore, qualities not inherent in the data, but external impositions to sift, categorise, and 'clean' data from extraneous 'noise'. Data that has been transformed into information has 'useless' elements removed and is given context and significance according to an external yardstick of 'usefulness'. To follow the earlier example, linking temperature readings in different rooms at different times, with occupancy readings and fluctuating electricity prices could be used by a 'smart' thermostat to make 'intelligent' heating choices.

Knowledge is to make sense of information, being aware of the limitations of the underlying data and theoretical models used to classify it, being able to place that information into a wider context of meaning, purpose, and dynamic interactions, involving experience, prediction, and the malleability of both purpose and model. Knowledge refers to the ability to understand a phenomenon, theoretically or practically, and to use such understanding for a deliberate purpose. It can be defined as 'justified true belief'.[13] This process complements available information with inferences from past experience and intuition, and responds to feedback, including sensory, cognitive, and evaluative.

Intelligence refers to the ability to 'function appropriately and with foresight', thus AI presumes that the act of thinking that turns (sensory) data into information and then into knowledge, and finally into purposeful action is not unique to humans or animals. It posits that the underlying computational process is formally deducible, can be scientifically studied and replicated in a digital computer. Once this is achieved, all the inherent advantages of the computer come to bear: speed, objectivity (absence of bias, emotion, preconceptions, etc.), scalability, permanent operation, etc. In the national security field, some have compared this promise to the mythical figure of the *Centaur*, who combined the intelligence of man with the speed and strength of the horse.[14]

The development of the Internet concerned the distribution of data and information between human and machine users.[15] AI, by contrast, does not primarily refer to the transmission of raw or processed data, the exchange of ideas, or the remote control of machinery (Internet of things, military command and control, etc.), but the ability to detect patterns in data, process data into information, and classify that information in order to predict outcomes and make decisions. *Darrell M. Allen* and *John R. West* suggest three differentiating characteristics of such systems: intentionality, intelligence, and adaptability.[16]

The Internet has already transformed our lives, but the enormous changes portended by AI are just beginning to dawn on us. The difficulty of predicting that change, however, should not serve as an excuse for what *James Baker* deemed 'a dangerous nonchalance' on behalf of decision-makers tasked with managing this transformation.[17] Responsible management of

[12] On this point, see generally E Derman, *Models Behaving Badly, Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life* (2011); I Hacking, *Representing and Intervening, Introductory Topics in the Philosophy of Natural Science* (1983).

[13] J Jenkins, M Steup, 'The Analysis of Knowledge' in E N Zalta (ed) *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.) https://plato.stanford.edu/entries/knowledge-analysis/.

[14] JE Baker, *The Centaur's Dilemma – National Security Law for the Coming AI Revolution* (2021) (hereafter Baker, The Centaur's Dilemma).

[15] See generally, BM Leiner and others, *Brief History of the Internet* (1997) (hereafter Leiner and others, Brief History of the Internet); M Waldrop, 'DARPA and the Internet Revolution' (DARPA, 2015) www.darpa.mil/about-us/timeline/modern-internet (hereafter Waldrop, 'DARPA and the Internet Revolution').

[16] DM West and JR Allen, 'How Artificial Intelligence Is Transforming the World' in *A Blueprint for the Future of AI* (Brookings, 24 April 2018) www.brookings.edu/series/a-blueprint-for-the-future-of-ai/.

[17] See note 3.

national security requires an adequate and realistic assessment of the threats and opportunities presented by new technological developments, especially their effect on the relative balance of power and on global public goods, such as the mitigation of catastrophic risks, arms races, and societal dislocations. In modern administrative states, such management is inevitably done through law, both nationally and internationally.[18]

In this chapter, I will begin by contrasting the challenge posed by AI to the related but distinct emergence of the cyber domain. I then outline six distinct implications for national security: doomsday scenarios, autonomous weapons, existing military capabilities, reconnaissance, economics, and foreign relations. Legal scholarship often proposes new regulation when faced with novel societal or technological challenges. But it appears unlikely that national actors will forego the potential advantages offered by a highly dynamic field through self-restraint by international convention. Still, even if outright bans and arms control-like arrangements are unlikely, the law serves three important functions when dealing with novel challenges: first, as the repository of essential values guiding action; second, offering essential procedural guidance; and third, by establishing authority, institutional mandates, and necessary boundaries for oversight and accountability.

II. CYBERSPACE AND AI

The purpose of this sub-section is not to outline the large literature applying the principles of general international law, and especially the law of armed conflict, to cyber operations. Rather, it seeks to highlight the distinctive elements of the global communication infrastructure, especially how AI is distinct from some of the regulatory and operational[19] challenges that characterise cybersecurity.[20] The mental image conjured by early utopian thinkers and adopted later by realist and military policy-makers rests on the geographical metaphor of 'cyberspace' as a non-corporeal place of opportunity and risk.[21] This place needs to be defended and thus constitutes an appropriate area of military operations.

As technical barriers eventually fell, the complexity of the network receded behind increasingly sophisticated but simple to operate graphical user-interfaces, making networked information-sharing first a mainstream, and eventually a ubiquitous phenomenon, affecting

[18] The literature on the administrative state is too copious to list, disparate discussions that helped guide my own thinking on this matter include S Cassese, 'Administrative Law without the State? The Challenge of Global Regulation' (2005) 37 *NYU Journal of International Law & Politics* 663; PD Feaver, 'The Civil-Military Problematique: Huntington, Janowitz, and the Question of Civilian Control' (1996) 23 *Armed Forces & Society* 149; SJ Kaufman, 'The Fragmentation and Consolidation of International Systems' (1997) 51 IO 173; A Chayes, 'An Inquiry into the Workings of Arms Control Agreements' (1972) 85 *Harvard Law Review* 905; AH Chayes and A Chayes, 'From Law Enforcement to Dispute Settlement: A New Approach to Arms Control Verification and Compliance' (1990) 14 *IS* 147.

[19] Good overviews can be found in GD Brown, 'Commentary on the Law of Cyber Operations and the DoD Law of War Manual' in MA Newton (ed), *The United States Department of Defense Law of War Manual* (2019); WH Boothby, 'Cyber Capabilities' in WH Boothby (ed), *New Technologies and the Law in War and Peace* (2018) (hereafter Boothby, 'Cyber Capabilities'); MN Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2017) 401 *et seq*.; JC Woltag, *Cyber Warfare: Military Cross-Border Computer Network Operations under International Law* (2014) (hereafter Woltag, *Cyber Warfare*); C Droege, 'Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians' (2012) 94 *Int'l Rev of the Red Cross* 533.

[20] With respect to cyber warfare, see also Chayes, 'Cyber Attacks and Cyber Warfare' (n 7); M Finnemore and DB Hollis, 'Constructing Norms for Global Cybersecurity' (2016) 110 *AJIL* 425. Regarding the specific impact of AI, see Baker, *The Centaur's Dilemma* (n 14).

[21] See further J Branch, 'What's in a Name? Metaphors and Cybersecurity' (2021) 75 *IO* 39 (hereafter Branch, 'Metaphors and Cybersecurity').

almost all aspects of human life almost everywhere. This has led to an exponential increase in the availability of information, much of it of a sensitive nature, often voluntarily relinquished. This has created a three-pronged challenge: data protection, information management, and network security.[22]

Much early civilian, especially academic, thinking focused on the dynamic relationship between technology and culture, stressing the emergence of a new, virtual habitat: 'A new universe, a parallel universe created and sustained by the world's computers and communication lines.'[23] But as the novelty wore off while its importance grew, the Internet became 're-territorialised' as nation-states asserted their jurisdiction, including in the hybrid, multi-stakeholder regulatory fora that had developed initially under American governmental patronage.[24] Perhaps more importantly, this non-corporeal realm created by connected computers, came to be seen not as a parallel universe following its own logic and laws, but as an extension of existing jurisdictions and organisational mandates:

> Although it is a man-made domain, cyberspace is now as relevant a domain for DoD [Department of Defence] activities as the naturally occurring domains of land, sea, air, and space. Though the networks and systems that make up cyberspace are man-made, often privately owned, and primarily civilian in use, treating cyberspace as a domain is a critical organizing concept for DoD's national security missions. This allows DoD to organize, train, and equip for cyberspace as we do in air, land, maritime, and space to support national security interests.[25]

This is reflected in the United States (US) National Security Strategy, which observes: 'Cybersecurity threats represent one of the most serious national security, public safety, and economic challenges we face as a nation.'[26] Other countries treat the issue with similar seriousness.[27]

Common to the manner in which diverse nations envisage cybersecurity is the emphasis on information infrastructure, in other words, on the need to keep communication channels operational and protected from unwanted intrusion. This, however, is distinct from the specific challenge of AI, which concerns the creation of actionable knowledge by a machine.

The initial ideas that led to the creation of the Internet sought to solve two distinct problems: the civilian desire to use expensive time-share computing capacity at academic facilities more efficiently by distributing tasks, and the military need to establish secure command and control

[22] See generally GD Solis, *The Law of Armed Conflict. International Humanitarian Law in War* (2016) 673–709 (hereafter Solis, *The Law of Armed Conflict*).

[23] ML Benedikt, *Cyberspace: First Steps* (1991) 1.

[24] See *inter alia* U Kohl, *The Net and the Nation State: Multidisciplinary Perspectives on Internet Governance* (2017) (hereafter Kohl, *The Net and the Nation State*); J Nocetti, 'Contest and Conquest: Russia and Global Internet Governance' (2015) 91 *Int'l Aff* 111; DeNardis, 'The Emerging Field of Internet Governance' (n 7); ML Mueller, *Networks and States: The Global Politics of Internet Governance* (2010).

[25] Department of Defence, 'Strategy for Operating in Cyberspace' (July 2011) https://csrc.nist.gov/CSRC/media/Projects/ISPAB/documents/DOD-Strategy-for-Operating-in-Cyberspace.pdf 5, referring to the 2010 Quadrennial Defence Review. Outer space has been an area of great power competition since the Sputnik satellite, but it has received added impetus in recent years with the creation of dedicated Space Commands in the US and other countries, see WJ Broad, 'How Space Became the Next 'Great Power' Contest between the US and China' *New York Times* (24 January 2021) www.nytimes.com/2021/01/24/us/politics/trump-biden-pentagon-space-missiles-satellite.html.

[26] Department of Defence, 'Strategy for Operating in Cyberspace' (July 2011) https://csrc.nist.gov/CSRC/media/Projects/ISPAB/documents/DOD-Strategy-for-Operating-in-Cyberspace.pdf 1, referring to the 2010 National Security Strategy. Similar language can be found in previous and subsequent national security strategies, both American and others, including the current 2021 interim one issued by the Biden Administration.

[27] E Afsah, 'Country Report Denmark' in M Kilching and C Sabine (eds), *Economic and Industrial Espionage in Germany and Europe* (2016).

connections between installations, especially to remote nuclear weapons facilities.[28] In both cases, it was discovered that existing circuit switched telephone connections were unreliable. The conceptional breakthrough consisted in the idea of package switched communication, which permitted existing physical networks to be joined non-hierarchically, permitting a non-hierarchical, decentralised architecture that is resilient, scalable, and open.[29]

The Internet is, therefore, not one network, but a set of protocols specifying data formats and rules of transmission, permitting local, physical networks to communicate along dynamically assigned pathways.[30] The technology, the opportunities, and the vulnerabilities it offered came to be condensed in the spatial analogy of cyberspace. This 'foundational metaphor' was politically consequential because the use of certain terminology implied, rather than stated outright, particular understandings of complex issues at the expense of others, thus shaping policy debates and outcomes.[31] Denounced later by himself as merely an 'effective buzzword' chosen because 'it seemed evocative and essentially meaningless', the definition offered by *William Gibson* highlights the problematic yet appealing character of this spatial analogy: 'Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation . . . A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding.'[32] The term combined the non-physical nature of a world being dynamically created by its denizens in their collective imagination, but relying behind the graphical user-interface on a complex physical infrastructure.[33] The advantages of open communications have eventually led military and

[28] The need to ensure reliable communication after sustaining a devastating first strike was a key ingredient of credible nuclear deterrence. The Soviet 'Dead Hand' system (*Mertvaya Ruka*, officially: *Systema Perimetr*) was an alternative, 'fail-deadly' method of solving that practical problem: meant as a backup to the *Kazbek* communication system, *Perimetr* was to fully automatically trigger nuclear retaliation if it detected an attack, even if command structures and human personnel had been destroyed. US Defence Intelligence Agency, 'Russia Military Power: Building a Military to Support Great Power Aspirations' (2017) https://www.hsdl.org/?view&did=801968 26–28; N Thompson, 'Inside the Apocalyptic Soviet Doomsday Machine' Wired (21 September 2009) www.wired.com/2009/09/mf-deadhand/; WJ Broad, 'Russia Has 'Doomsday' Machine, US Expert Says' *New York Times* (8 October 1993) www.nytimes.com/1993/10/08/world/russia-has-doomsday-machine-us-expert-says.html.

[29] This means that data to be transmitted will be split into several packets, based on various criteria including size. The packets will be sent independently from each other, usually along different pathways, and re-assembled at the destination. They contain the actual data to be sent, destination and source address, and other information necessary for reliable transmission. The idea was simultaneously but independently developed at MIT in Cambridge, Massachusetts (1961–1967), RAND in Santa Monica, California (1962–1965) and the British National Physical Laboratory (NPL) in London (1964–1967). This genesis is well described by several of its key protagonists themselves in Leiner and others, *Brief History of the Internet* (n 15); Waldrop, 'DARPA and the Internet Revolution' (n 15).

[30] This reliance on a conceptional, rather than physical architecture is reflected in the definition laid down in US law: 'The term "Internet" means collectively the myriad of computer and telecommunications facilities, including equipment and operating software, which comprise the interconnected world-wide network of networks that employ the Transmission Control Protocol/Internet Protocol, or any predecessor or successor protocols to such protocol, to communicate information of all kinds by wire or radio.'15 USC § 6501(6), www.law.cornell.edu/uscode/text/15/6501#6.

See also Woltag, *Cyber Warfare* (n 19) 9.

[31] Branch, 'Metaphors and Cybersecurity' (n 21).

[32] W Gibson, *Neuromancer* (1984) 69, emphasis added. Gibson makes the disparaging remarks about his term in the documentary film M Neale, 'No Maps for these Territories' (2000).

[33] 'Gibson's networked artificial environment anticipated the globally internetworked technoculture (and its surveillance) in which we now find ourselves. The term has gone on to revolutionize popular culture and popular science, heralding the power and ubiquity of the information age we now regard as common as iPhones. Since its invention, 'cyberspace' has come to represent everything from computers and information technology to the Internet and "consensual hallucinations" as different as The Matrix, Total Information Awareness, and reality TV.' March 17, 1948: W Gibson, 'Father of Cyberspace' *Wired* (16 March 2009) www.wired.com/2009/03/march-17-1948-william-gibson-father-of-cyberspace-2/.

civilian installations in all nations to become accessible through the Internet, creating unique vulnerabilities due to opportunity costs of communication disruption, physical damage to installations, and interruptions of critical public goods like water or electricity.[34] What the American military defines as its key challenge in this area applies likewise to most other nations:

> US and international businesses trade goods and services in cyberspace, moving assets across the globe in seconds. In addition to facilitating trade in other sectors, cyberspace is itself a key sector of the global economy. Cyberspace has become an incubator for new forms of entrepreneurship, advances in technology, the spread of free speech, and new social networks that drive our economy and reflect our principles. The security and effective operation of US critical infrastructure – including energy, banking and finance, transportation, communication, and the Defense Industrial Base – rely on cyberspace, industrial control systems, and information technology that may be vulnerable to disruption or exploitation.[35]

Some have questioned the definitional appropriation of 'cyberspace' as a 'domain' for military action through 'linguistic and ideational factors [which] are largely overlooked by the prevailing approach to cybersecurity in IR [international relations], which has productively emphasized technical and strategic aspects' at the expense of alternative ways of thinking about security in this field.[36] Without prejudice to the theoretical contributions such investigations could make to political science and international relations,[37] the legal regulation of defensive and offensive networked operations has, perhaps after a period of initial confusion,[38] found traditional concepts to be quite adequate, perhaps because the spatial analogy facilitates the application of existing legal concepts.

The central challenges posed by the increasing and unavoidable dependence on open-architecture communication are both civilian and military. They concern primarily three distinct but related operational tasks: prevent interruptions to the flow of information, especially financial transactions; prevent disruptions to critical command and control of civilian and military infrastructure, especially energy, water, and nuclear installations; and prevent unauthorised access to trade and military secrets.[39] These vulnerabilities have, of course, corresponding opportunities for obtaining strategic information, striking at long distance while maintaining 'plausible deniability',[40] and establishing credible deterrence.[41] Again, how the American military describes its own mandate applies in equal measure to other nations, not least its chief competitors Russia and China:

[34] DE Sanger, 'China Appears to Warn India: Push Too Hard and the Lights Could Go Out' *New York Times* (28 February 2021) www.nytimes.com/2021/02/28/us/politics/china-india-hacking-electricity.html.

[35] US Department of Defence, 'Strategy for Operating in Cyberspace' (July 2011) 1, https://csrc.nist.gov/CSRC/media/Projects/ISPAB/documents/DOD-Strategy-for-Operating-in-Cyberspace.pdf.

[36] Branch, 'Metaphors and Cybersecurity' (n 21) 41.

[37] See for instance M Finnemore and DB Hollis, 'Constructing Norms for Global Cybersecurity' (2016) 110 *AJIL* 425.

[38] A Chayes, 'Implications for Civil-Military Relations in Cyber Attacks and Cyber Warfare' in A Chayes (ed), *Borderless Wars: Civil Military Disorder and Legal Uncertainty* (2015).

[39] Politiets Efterretningstjeneste, 'Trusler mod Danmark: Spionage' (2015), https://pet.dk/spionage; JUO Nielsen, 'Erhvervshemmelighedsværnet i Norden og EU' (2014) *Erhvervsjuridisk Tidsskrift* 1.

[40] See further L Arimatsu, 'The Law of State Responsibility in Relation to Border Crossings: An Ignored Legal Paradigm' (2013) 89 *Int'l L Stud* 21; P Margulies, 'Networks in Non-International Armed Conflicts: Crossing Borders and Defining "Organized Armed Groups"' (2013) 89 *Int'l L Stud* 54.

[41] Y Benkler, 'Degrees of Freedom, Dimensions of Power' (2016) *Daedalus* 18 (hereafter Benkler, 'Degrees of Freedom'). Unlike in classical military spheres, it is important to note that in the cyber-domain effective repulsion and deterrence does not necessarily have to be assumed by the military, see Forsvarsministeriet, 'Center for Cybersikkerhed' (18 September 2020) https://www.fmn.dk/da/arbejdsomraader/cybersikkerhed/center-for-cybersikkerhed/.

> American prosperity, liberty, and security depend upon open and reliable access to information. The Internet empowers us and enriches our lives by providing ever-greater access to new knowledge, businesses, and services. Computers and network technologies underpin US military warfighting superiority by enabling the Joint Force to gain the information advantage, strike at long distance, and exercise global command and control.
>
> The arrival of the digital age has also created challenges for the Department of Defense (DoD) and the Nation. The open, transnational, and decentralized nature of the Internet that we seek to protect creates significant vulnerabilities. Competitors deterred from engaging the US and our allies in an armed conflict are using cyberspace operations to steal our technology, disrupt our government and commerce, challenge our democratic processes, and threaten our critical infrastructure.[42]

Crucially important as these vulnerabilities and opportunities are for national security, defensive and offensive operations occurring on transnational communication networks raise important regulatory questions,[43] including the applicability of the law of armed conflict to so-called cyber-operations.[44] *Yoram Dinstein* dismisses the need for a revolution in the law of armed conflict necessitated by the advent of cyber warfare: 'this is by no means the first time in the history of LOAC that the introduction of a new weapon has created the misleading impression that great legal transmutations are afoot. Let me remind you of what happened upon the introduction of another new weapon, viz., the submarine.'[45] *Dinstein* recounts how the introduction of the submarine in World War I led to frantic calls for international legal regulation. But instead of comprehensive new conventional law, states eventually found the mere restatement that existing rules must also be observed by submarines sufficient. He concludes that were an international convention on cyber warfare to be concluded today, 'it would similarly stipulate in an anodyne fashion that the general rules of LOAC must be conformed with.'[46] *Gary Solis* likewise opens the requisite chapter in his magisterial textbook by stating categorically: 'This discussion is out of date. Cyber warfare policy and strategies evolve so rapidly that is difficult to stay current.' But what is changing are technologies, policies, and strategies, not the law: 'Actually, cyber warfare issues may be resolved in terms of traditional law of war concepts, although there is scant demonstration of its application because, so far, instances of actual cyber warfare have been unusual. Although cyber questions are many, the law of war offers as many answers.'[47] Concrete answers will depend on facts that are difficult to ascertain, due to inherent technical difficulties to forensic analysis in an extremely complex, deliberately heterogeneous network composed of a multitude of actors, both private and public, benign and malign. Legal assessments likewise rely on definitional disputes and normative interpretations that reflect shifting, often short-term, policies and strategies. Given vastly divergent national interests and capabilities, no uniform international understanding, let alone treaty regulation has emerged.[48]

In sum, while AI relies heavily on the same technical infrastructure of an open, global information network, its utilisation in the national security field poses distinct operational and

[42] Department of Defence, 'Cyber Strategy 2018 – Summary' (2018) https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF 1.

[43] Kohl, *The Net and the Nation State* (n 24); DeNardis, 'The Emerging Field of Internet Governance' (n 7).

[44] Boothby, 'Cyber Capabilities' (n 19); WH Boothby, 'Methods and Means of Cyber Warfare' (2013) 89 *Int'l L Stud* 387.

[45] RN Chesney, 'Computer Network Operations and US Domestic Law: An Overview' (2013) 87 *International Law Studies* 218, 286.

[46] Ibid, 287.

[47] Solis, *The Law of Armed Conflict* (n 21) 673.

[48] But note the highly representative Tallinn Manual, see W Heintschel von Heinegg, 'Chapter 1: The Tallinn Manual and International Cyber Security Law' (2012) 15 *YBIHL* 3.

legal challenges not fully encompassed by the law of 'cyber warfare'.[49] That area of law presents the lawyer primarily with the challenge of applying traditional legal concepts to novel technical situations, especially the evidentiary challenges of defining and determining an armed attack, establishing attribution, the scope of the right to self-defence and proportionality, as well as thorny questions of the treatment of non-state or quasi-state actors, the classification of conflicts, and not least the threshold of the 'use of force'.[50] AI sharpens many of the same regulatory *conundra*, while creating novel operational risks and opportunities.[51]

III. CATASTROPHIC RISK: DOOMSDAY MACHINES

In the latest instalment of the popular Star Wars movie franchise, there is a key scene where the capabilities of truly terrible robotic fighting machines are presented. The franchise's new hero, the eponymous *Mandalorian*, manages only with considerable difficulty to defeat but one of these robots, of which, however, an entire battalion is waiting in the wings. The designers of the series have been praised for giving audiences 'finally an interesting stormtrooper', that is a machine capable of instilling fear and respect in the viewer.[52]

Whatever the cineastic value of these stormtroopers, in a remarkable coincidence a real robotics company simultaneously released a promotional video of actual robots that made these supposedly frightening machines set in a far distant future look like crude, unsophisticated toys. The dance video released by Boston Dynamics in early 2021 to show off several of its tactical robots jumping, dancing, pirouetting elegantly to music put everything Hollywood had come up with to shame: these were no prototypes, but robots that had already been deployed to police departments[53] and the military,[54] doing things that one previously could only have imagined in computer generated imagery.[55] Impressive and fearsome as these images are, these robots do exhibit motional 'intelligence' in the sense that they are able to make sense of their surroundings and act purposefully in it, but they are hardly able to replicate, let alone compete with human action, yet.

The impressive, even elegant capabilities showcased by these robots show that AI has made dramatic strides in recent years, bringing to mind ominous fears. In an early paper written in 1965, one of the British Bletchley Park cryptographers, the pioneering computer scientist and friend of *Alan Turing*, *Irving John 'Jack' Good* warned that an 'ultra-intelligent machine' would be built in the near future that could prove to be mankind's 'last invention' because it would lead to an 'intelligence explosion', that is an exponential increase in self-generating machine

[49] An excellent overview is provided by Solis, *The Law of Armed Conflict* (n 22) 673–709.

[50] MN Schmitt, 'The Law of Cyber Warfare: Quo Vadis?' (2014) 25 *Stanford Law & Policy Review* 269, 279.

[51] See in Baker, *The Centaur's Dilemma* (n 14) 69–94.

[52] J Hellerman, '"The Mandalorian" Finally Gives Us an Interesting Stormtrooper' (*No Film School Blog*, 18 December 2020) https://nofilmschool.com/storm-troopers-dumb.

[53] A Olla, 'A Dystopian Robo-Dog Now Patrols New York City. That's the Last Thing We Need' *The Guardian* (2 March 2021) www.theguardian.com/commentisfree/2021/mar/02/nypd-police-robodog-patrols.

[54] The humanoid Russian FEDOR tactical robot has already been deployed to the International Space Station, L Grush, 'Russia's Humanoid Robot Skybot Is on Its Way Home After a Two-Week Stay in Space' (*The Verge*, 6 September 2019) www.theverge.com/2019/9/6/20852602/russia-skybot-fedor-robot-international-space-station-soyuz.

[55] The video carried a note that these were not digital images but real footage of actual robots. See also E Ackerman, 'How Boston Dynamics Taught Its Robots to Dance' (*IEEE Spectrum*, 7 January 2021) https://spectrum.ieee.org/automaton/robotics/humanoids/how-boston-dynamics-taught-its-robots-to-dance; B Gilbert, 'Watch a Rare Video of Robots Jumping and Dancing Inside One of America's Leading Robotics Firms' *Business Insider* (29 March 2021) www.businessinsider.com/video-robots-jumping-and-dancing-inside-boston-dynamics-2021-3.

intelligence.[56] While highly agile tactical robots conjure tropes of dangerous machines enslaving humanity, the potential risk posed by the emergence of super-intelligence is unlikely to take either humanoid form or motive but constitutes both incredible opportunity and existential risk, as *Good* pointed out half a century ago:

> The survival of man depends on the early construction of an ultra-intelligent machine. ... Let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus, the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.[57]

Good would have been pleased to learn that both the promise and premonition of AI are no longer the preserve of science fiction, but taken seriously at the highest level of political decision-making. In a well-reported speech, President Vladimir Putin of Russia declared in 2017 that leadership in AI: 'is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world.'[58] Very similar statements guide official policy in all great powers, raising the spectre of what has been termed an 'arms race' in AI,[59] as a result of which 'super-intelligent' machines (i.e. those with capabilities higher than humans across the board), might endanger mankind.[60]

It is interesting to note that the tone of the debate has changed significantly. Writing in a popular scientific magazine in 2013, *Seth Baum* asked rhetorically whether his readers should even take the topic seriously: 'After all, it is essentially never in the news, and most AI researchers don't even worry. (AGI today is a small branch of the broader AI field.) It's easy to imagine this to be a fringe issue only taken seriously by a few gullible eccentrics.'[61] Today, these statements are no longer true. As Artificial General Intelligence, and thus the prospect of super-intelligence, is becoming a prominent research field, worrying about its eventual security implications is no longer the preserve of 'a few gullible eccentrics'. *Baum* correctly predicted that the relative lack of public and elite attention did not mean that the issue was unimportant.

Comparing it to the issue of climate change that likewise took several decades to evolve from a specialist concern to an all-consuming danger, he predicted that the trend was clear that given the exponential development of technology, the issue would soon become headline news.

[56] IJ Good, 'Speculations Concerning the First Ultraintelligent Machine' (1966) 6 *Advances in Computers* 31.

[57] Ibid, 31, 33, references omitted.

[58] J Vincent, 'Putin says the nation that leads in AI "will be the ruler of the world"', The Verge (4 September 2017) https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.

[59] The comprehensive study commissioned by the European Parliament on this topic lists existential risk only as the last item of twelve 'ethical harms and concerns' currently tackled by national and international regulatory efforts; E Bird and others, 'The Ethics of Artificial Intelligence: Issues and Initiatives' (*European Parliament*, March 2020) www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf> 42–43 (hereafter Bird and others, 'The Ethics of Artificial Intelligence').

[60] See also the section on 'Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)' in M Bourgon and R Mallah, 'Ethically Aligned Design – A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, (1st ed.)' (*IEEE*, 2019) https://ethicsinaction.ieee.org (hereafter Bourgon and Mallah, 'Ethically Aligned Design').

[61] S Baum, 'Our Final Invention: Is AI the Defining Issue for Humanity?' *Scientific American* (11 October 2013) https://blogs.scientificamerican.com/guest-blog/our-final-invention-is-ai-the-defining-issue-for-humanity/.

The same point was made roughly at the same time by the co-founder of the Centre for the Study of Existential Risk (CSER) at the University of Cambridge, *Huw Price*. Summing up the challenge accurately, Price acknowledged that some of these concerns might seem far-fetched, the stuff of science fiction, which is exactly part of the problem:

> The basic philosophy is that we should be taking seriously the fact that we are getting to the point where our technologies have the potential to threaten our own existence – in a way that they simply haven't up to now, in human history. We should be investing a little of our intellectual resources in shifting some probability from bad outcomes to good ones. To the extent – presently poorly understood – that there are significant risks, it's an additional danger if they remain for these sociological reasons outside the scope of 'serious' investigation.[62]

There are two basic options: either to design safe AI with appropriate standards of transparency and ethical grounding as inherent design features, or not to design dangerous AI.[63] Given the attendant opportunities and the competitive international and commercial landscape, this latter option remains unattainable. Consequently, there has been much scientific thinking on devising ethical standards to guide responsible further technological development.[64] International legal regulation, in contrast, has so far proven elusive, and national efforts remain embryonic.[65]

Some serious thinkers and entrepreneurs argue that the development of super-intelligence must be abandoned due to inherent, incalculable, and existential risks.[66] Prudence would indicate that even a remote risk of a catastrophic outcome should keep all of us vigilant. Whatever the merits of these assessments, it appears unlikely that an international ban of such research is likely. Moreover, as *Ryan Calo* and others have pointed out, there is a real opportunity cost in focusing too much on such remote but highly imaginative risks.[67]

While the risks of artificial super-intelligence, which is defined as machine intelligence that surpasses the brightest human minds, are still remote, they are real and may quickly threaten human existence by design or indifference. Likewise, general AI or human-level machine intelligence remains largely aspirational, referring to machines that can emulate human beings at a range of tasks, switching fluidly between them, training themselves on data and their own past performance, and re-writing their operating code. In contrast, concrete policy and regulatory challenges need to be addressed now as a result of the exponential development of the less fearsome but concrete narrow AI, defined as machines that are as good or better than humans at particular tasks, such as interpreting x-ray or satellite images.

These more mundane systems are already operational and rapidly increase in importance, especially in the military field. Here, perhaps even more than in purely civilian domains, *Pedro Domingos*' often quoted adage seems fitting: 'People worry that computers will get too smart and

[62] F Lewsey, 'Humanity's Last Invention and Our Uncertain Future' (*University of Cambridge*, 25 November 2012) www.cam.ac.uk/research/news/humanitys-last-invention-and-our-uncertain-future.

[63] To some extent, this debate is already moot because automated strategic nuclear defence systems have existed – and likely remain operational – in both Russia and the United States, see n 27.

[64] The evolving scientific, industry, and governmental consensus about the principles necessary to ensure responsible and safe AI have been outlined *inter alia* in Bourgon and Mallah, 'Ethically Aligned Design' (n 60): 'Asilomar Principles on Intelligent Machines and Smart Policies – Research Issues, Ethics and Values, Longer-Term Values' (*Future of Life Institute*, 2017) futureoflife.org/ai-principles.

[65] For an overview of national efforts, see Bird and others, 'The Ethics of Artificial Intelligence' (n 59).

[66] For an approving summary of these arguments, see J Barratt, *Our Final Invention* (2013).

[67] R Calo, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51 *University of California Davis Law Review* 399–435 (hereafter Calo, 'Artificial Intelligence Policy').

take over the world, but the real problem is that they're too stupid and they've already taken over the world.'[68] Without belittling the risk of artificial general or super-intelligence, *Calo* is thus correct to stress that focusing too much attention on this remote risk will reduce necessary attention from pressing societal needs and thereby risk 'an AI Policy Winter' in which necessary regulation limps behind rapid technical development.[69]

IV. AUTONOMOUS WEAPONS SYSTEM

Automated weapons have been in use for a long time; how long depends largely on the degree of automation informing one's definition. A broad definition of a robot, under which we can subsume autonomous weapons systems, is a physical system that senses, processes, and acts upon the world. We can thus differentiate between 'disembodied AI' which collects, processes, and outputs data and information, but whose effect in the physical world is mediated; and robotics which leverage AI to itself physically act upon the world.[70]

In order to ascertain the likely impact of AI on autonomous weapons systems, it is helpful to conceive of them and the regulatory challenges they pose as a spectrum of capabilities rather than sharply differentiated categories, with booby traps and mines on one end; improvised explosive devices (IEDs), torpedoes, and self-guided rockets somewhere in the middle; drones and loitering munition further towards the other end; and automated air defence and strategic nuclear control systems at or beyond the other polar end. It appears that two qualitative elements are crucial: the degree of processing undertaken by the system,[71] and the amount of human involvement before the system acts.[72]

It follows that the definition of 'autonomous' is not clear-cut, nor is it likely to become so. Analytically, one can distinguish four distinct levels of autonomy: human operated, human delegated, human supervised, and fully autonomous.[73] These classifications, however, erroneously 'imply that there are discrete levels of intelligence and autonomous systems',[74] downplaying the importance of human–machine collaboration.[75] Many militaries, most prominently that of the US, insist that a human operator must remain involved, including 'fail safe' security precautions:

[68] P Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (2015) 286.

[69] Calo, 'Artificial Intelligence Policy' (n 67) 435.

[70] Calo, 'Artificial Intelligence Policy' (n 67) 407. Calo argues that the respective legal assessment is likely to be different; see also HY Liu, 'Refining Responsibility: Differentiating Two Types of Responsibility Issues Raised by Autonomous Weapons Systems' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016) (hereafter Liu, 'Refining Responsibility').

[71] See further HY Liu, 'Categorization and Legality of Autonomous and Remote Weapons Systems' (2012) 94 *Int'l Rev of the Red Cross* 627; G Sartor and O Andrea, 'The Autonomy of Technological Systems and Responsibilities for their Use' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016).

[72] See further N Sharkey, 'Staying in the Loop: Human Supervisory Control of Weapons' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016) (hereafter Sharkey, 'Staying in the Loop'); GS Corn, 'Autonomous Weapons Systems: Managing the Inevitability of "Taking the Man Out of the Loop"' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016) (hereafter Corn, 'Autonomous Weapons Systems'); D Saxon, 'A Human Touch: Autonomous Weapons, DoD Directive 3000.09 and the Interpretation of "Appropriate Levels of Human Judgment over the Use of Force"' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016) (hereafter Saxon, 'A Human Touch').

[73] G Galdorisi, 'Keeping Humans in the Loop' (2015) 141/2/1,344 US Naval Institute Proceedings 36, 38.

[74] Department of Defense, Defense Science Board, 'Task Force Report: The Role of Autonomy in DoD Systems' (*US Department of Defense*, July 2012) 4 https://fas.org/irp/agency/dod/dsb/autonomy.pdf.

[75] G Galdorisi, 'Keeping Humans in the Loop' (2015) 141/2/1,344 US Naval Institute Proceedings 36; Sharkey, 'Staying in the Loop' (n 72).

> Semi-autonomous weapons systems that are onboard or integrated with unmanned platforms must be designed such that, *in the event of degraded or lost communications*, the system does *not autonomously select and engage* individual targets or specific target groups that have not been previously selected by an authorized human operator. It is DoD policy that . . . autonomous and semi-autonomous weapons systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment of the use of force.[76]

In contrast to the assumptions underlying the discussion in the previous section, even fully autonomous systems currently always involve a human being who 'makes, approves, or overrides a fire/don't fire decision'.[77] Furthermore, such systems have been designed by humans, who have programmed them within specified parameters, which include the need to observe the existing law of armed conflict.[78] These systems are deployed into battle by human operators and their commanders,[79] who thus carry command responsibility,[80] including the possible application of strict liability standards known from civil law.[81]

Given the apparent military benefits of increased automation and an extremely dynamic, easily transferable civilian field, outright bans of autonomous weapon systems, robotics, and unmanned vehicles appear 'insupportable as a matter of law, policy, and operational good sense'.[82] To be sure, some claim that the principles of distinction, proportionality, military necessity, and the avoidance of unnecessary suffering, which form the basis of the law of armed conflict,[83] in conjunction with general human rights law,[84] somehow impose a 'duty upon individuals and states in peacetime, as well as combatants, military organizations, and states in armed conflict situations, not to delegate to a machine or automated process the authority or capability to initiate the use of lethal force independently of human determinations of its moral and legal legitimacy in each and every case.'[85] Without restating the copious literature on this topic, it is respectfully suggested that such a duty for human determination cannot be found in existing international, and only occasionally in national,[86] law. *Solis*' textbook begins discussing the war crime liability of autonomous weapons by stating the

[76] Directive 3000.09: Autonomy in Weapons Systems, Unmanned Systems Integrated Roadmap, FY 2013–2038, US Department of Defence, Washington D.C. (21 November 2012); Solis, *The Law of Armed Conflict* (n 22) 537, my emphasis.

[77] Solis, *The Law of Armed Conflict* (n 22) 537.

[78] P Asaro, 'On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making' (2012) 94 *Int'l Rev of the Red Cross* 687, 691 (hereafetr Asaro, 'On Banning Autonomous Weapon Systems').

[79] MN Schmitt and JS Thurnher, "Out of the Loop': Autonomous Weapons Systems and the Law of Armed Conflict' (2013) 4 *Harvard National Security Journal* 231, 235 (hereafter Schmitt and Thurnher, 'Out of the Loop').

[80] Sharkey, 'Staying in the Loop' (n 72); Liu, 'Refining Responsibility' (n 70).

[81] Calo, 'Artificial Intelligence Policy' (n 67) 418; R Calo, 'Robotics and the Lessons of Cyberlaw' (2015) 103 *California Law Review* 513, 538–545 (hereafter Calo, 'Robotics and the Lessons of Cyberlaw').

[82] Schmitt and Thurnher, 'Out of the Loop' (n 79) 233.

[83] Solis, *The Law of Armed Conflict* (n 22) 268–327, 539–541, 551–552.

[84] M Milanovic, 'The Lost Origins of Lex Specialis: Rethinking the Relationship between Human Rights and International Humanitarian Law' in JD Ohlin (ed), *Theoretical Boundaries of Armed Conflict and Human Rights* (2015); G Pinzauti, 'Good Time for a Change: Recognizing Individuals' Rights under the Rules of International Humanitarian Law on the Conduct of Hostilities' in A Cassese (ed), *Realizing Utopia: The Future of International Law* (2012); T Meron, 'On the Inadequate Reach of Humanitarian and Human Rights Law and the Need for a New Instrument' (1983) 77 *AJIL* 589–606; T Meron, *Human Rights and Humanitarian Norms as Customary Law* (1991).

[85] Asaro, 'On Banning Autonomous Weapon Systems' (n 78) 687; E Lieblich and B Eyal, 'The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapons Systems Are Unlawful' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016).

[86] The American military, it is remembered, formally maintains that it is bound by such a duty, as a matter of internal policy. Whether this amounts to a legal obligation under domestic law remains a matter of some dispute; see further Department of Defence, *Directive 3000.09: Autonomy in Weapons Systems* (n 77); Saxon, 'A Human Touch' (n 72).

obvious counter-factual: 'Any lawful weapon can be employed unlawfully.' He proceeds to devise a number of hypothetical scenarios in which autonomous weapons could indeed be used or deliberately designed unlawfully, to conclude:

> The likelihood of an autonomous weapon system being unlawful in and of itself is very remote; it would not meet Article 36 testing requirements and thus would not be put into use. And the foregoing four scenarios involving possible unlawful acts by operators or manufacturers are so unlikely, so phantasmagorical, that they are easily lampooned. ... While acts such as described in the four scenarios are unlikely, they are possible.[87]

As stated, Article 36 of the 1977 Additional Protocol I to the Geneva Conventions imposes on the contracting parties the obligation to determine prior to the deployment of any new weapon that it conforms with the existing law of armed conflict and 'any other rule of international law applicable'. For states developing new weapons, this obligation entails a continuous review process from conception and design, through its technological development and prototyping, to production and deployment.[88]

Given the complexity and rapid continuous development of autonomous weapons systems, especially those relying on increasingly sophisticated AI, such a legally mandatory review will have to be continuous, rigorous, and overcome inherent technical difficulties, given the large number of sub-systems from a large number of providers. Such complexity notwithstanding, autonomous weapons, including those relying on AI, are not unlawful in and of themselves.

In principle, the underlying ethical *conundra* and proportional balancing of competing values that need to inform responsible robotics generally,[89] need to inform the conception, design, deployment, and use of autonomous weapons system, whether or not powered by AI: 'I reject the idea that IHL [international humanitarian law] is inadequate to regulate autonomous weapons. ... However far we go into the future and no matter how artificial intelligence will work, there will always be a human being at the starting point ... This human being is bound by the law.'[90] The most likely use scenarios encompass so-called narrow AI where machines have already surpassed human capabilities. The superior ability to detect patterns in vast amounts of unstructured (sensory) data has for many years proven indispensable for certain advanced automated weapons systems. Anti-missile defence systems, like the American maritime Aegis and land-based Patriot, the Russian S300 and S400 or the Israeli 'Iron Dome', all rely on the collection and processing of large amounts of radar and similar sensor data, and the ability to respond independently and automatically. This has created unique vulnerabilities: their susceptibility to cyber-attacks 'blinding' them,[91] the dramatic shortening of warning and reaction time even where human operators remain 'in the loop',[92] and the possibility to render these

[87] Solis, *The Law of Armed Conflict* (n 22) 543.

[88] International Committee of the Red Cross (ICRC), *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare* (2006) 23.

[89] See generally Calo, 'Robotics and the Lessons of Cyberlaw' (n 81).

[90] See further M Sassòli, 'Autonomous Weapons and International Law: Advantages, Open Technical Questions and Legal Issues to be Clarified' (2014) 90 *International Law Studies* 308, 323; likewise, Schmitt and Thurnher, 'Out of the Loop' (n 79) 277.

[91] Note for instance the Israeli electronic disabling of Syria's expensive, Russian-made air defence system prior to their bombing of a half-constructed nuclear power reactor in 2007, discussed in Solis, *The Law of Armed Conflict* (n 22) 677.

[92] This has been the main excuse offered by the Captain of the US warship *Vincennes* for shooting down an Iranian civilian airliner in 1988. With AI, this problem is likely to become much more acute. For a discussion of the former, see ibid, 563–566. For the latter, see Baker, 'Artificial Intelligence and National Security Law' (n 3).

expensive, highly sophisticated systems economically unviable by targeting them with unconventional countermeasures, such as very cheap, fairly simple commercial drones.[93]

V. EXISTING MILITARY CAPABILITIES

Irrespective of the legal and ethical questions raised, AI is having a transformative effect on the operational and economic viability of many sophisticated weapons systems. The existing military technology perhaps most immediately affected by the rise of AI are unmanned vehicles of various kinds, so-called drones and 'loitering munitions'.[94] Currently relying on remote guidance by human operators or relatively 'dumb' automation, their importance and power is likely to increase enormously if combined with AI. Simultaneously, certain important legacy systems, for instance large surface ships such as aircraft carriers, can become vulnerable and perhaps obsolete due to neurally linked and (narrowly) artificially intelligent 'swarms' of very small robots.[95]

The ready availability of capable and affordable remotely operated vehicles, plus commercial satellite imagery and similar information sources has put long-range power-projection capabilities in the hands of a far larger group of state and non-state actors. This equalisation of relative power is further accelerated by new technology rendering existing weapon systems vulnerable or ineffective. Important examples include distributed, swarm-like attacks on ships or permeating expensive air defence systems with cheap, easily replaceable commercial drones.[96]

The recent war over Nagorno-Karabakh exposed some of these general vulnerabilities, not least the inability of both Armenia and Azerbaijan's short-range air defense (SHORAD) arsenals, which admittedly were limited in size and quality, to protect effectively against sophisticated drones. While major powers like the US, China, and Russia are developing and deploying their own drone countermeasures,[97] certain existing systems, for instance aircraft carriers, have become vulnerable. This portends potential realignments in relative power where large numbers of low-cost expendable machines can be used to overwhelm an otherwise superior adversary.[98]

There has been much academic speculation about the perceived novelty of drone technology and the suggested need to update existing legal regulations.[99] It needs to be stated from the outset that remotely piloted land-, air-, or sea-crafts have been used since the 1920s,[100] and thus

[93] Note for instance the successful use by Houthi militias in Yemen and by Hamas in Gaza of very cheap commercial drones as deliberate targets for very expensive Israeli, Emirati, and Saudi Patriot air defence systems; see A Kurth Cronin, *Power to the People: How Drones, Data and Dynamite Empower and Imperil Our Security* (2019) 213.

[94] These are sometimes called 'suicide drones.' For an excellent technical overview, see D Gettinger and HM Arthur, 'Loitering Munitions' (*CSD Bard*, 2017) https://dronecenter.bard.edu/files/2017/02/CSD-Loitering-Munitions.pdf.

[95] T McMullan, 'How Swarming Drones Will Change Warfare' (*BBC News*, 16 March 2019) www.bbc.com/news/technology-47555588 (hereafter McMullan, 'How Swarming Drones Will Change Warfare; SM Williams, 'Swarm Weapons: Demonstrating a Swarm Intelligent Algorithm for Parallel Attack' (2018) https://apps.dtic.mil/sti/pdfs/AD1071535.pdf (hereafter Williams, 'Swarm Weapons').

[96] R Martinage, 'Toward a New Offset Strategy – Exploiting US Long-Term Advantages to Restore US Global Power Projection Capability' (CSBA, 2014) https://csbaonline.org/uploads/documents/Offset-Strategy-Web.pdf 23–28 (hereafter Martinage, 'Toward a New Offset Strategy').

[97] S Shaikh and R Wes, 'The Air and Missile War in Nagorno-Karabakh: Lessons for the Future of Strike and Defense' (CSIC, 8 December 2020) www.csis.org/analysis/air-and-missile-war-nagorno-karabakh-lessons-future-strike-and-defense (hereafter Shaikh and Wes, 'Lessons for the Future of Strike and Defense').

[98] McMullan, 'How Swarming Drones Will Change Warfare' (n 95); Williams, 'Swarm Weapons' (n 95).

[99] For an overview, see PL Bergen and D Rothenberg (eds), *Drone Wars: Transforming Conflict, Law, and Policy* (2015) (hereafter Bergen and Rothenberg, Drone Wars).

[100] K Kakaes, 'From Orville Wright to September 11: What the History of Drone Technology Says about Its Future' in Bergen and Rothenberg, *Drone Wars: Transforming Conflict, Law, and Policy* (2015) (hereafter Kakaes, 'From Orville Wright to September 11').

cannot be considered either new or unanticipated by the existing law of armed conflict.[101] Likewise, it is difficult to draw a sharp technical distinction between certain drones and some self-guided missiles, which belong to a well-established area of military operations and regulation.[102]

The novelty lies less in the legal or ethical assessment, than in the operational challenge of the dispersal of a previously highly exclusive military capability. The US has twice before responded to such a loss of its superior competitive edge by embarking on an 'offset' strategy meant to avoid having to match capabilities, instead seeking to regain superiority through an asymmetric technological advantage.[103]

The 'First Offset' strategy successfully sought to counter Soviet conventional superiority through the development and deployment of, especially tactical, nuclear weapons.[104] The 'Second Offset' strategy was begun towards the end of the Vietnam War and reached its successful conclusion during the Iraq War of 1991. It meant to counter the quantitative equalisation of conventional assets, especially airpower, not by increasing the number of assets but their quality. Mustering American socio-economic advantages in technological sophistication, the key to the strategy was the development of previously unimaginable strike precision. As with any other military technology, it was anticipated that the opponent would eventually catch up, at some point neutralising this advantage. Given the economic near-collapse of the Soviet Union and its successor Russia, the slow rise of China, and the relative absence of other serious competitors, the technological superiority the US had achieved in precision strike capability surprisingly endured far longer than anticipated:

> Perhaps the most striking feature of the evolution of non-nuclear (or conventional) precision strike since the Cold War ended in 1991 has been what has not happened. In the early 1990s, there was growing anticipation that for major powers such as the United States and Russia, 'long-range precision strike' would become 'the dominant operational approach.' The rate at which

[101] For a good overview, see Solis, *The Law of Armed Conflict* (n 22) 545–554. The claim that the existing law of armed conflict is inadequate for the actual conflict at hand is probably as old as the truism that this body of law is 'always one war behind.' While there is some truth in the latter observation, the first is usually little more than exculpatory. Both discussions are as old as humanitarian law itself and it is unlikely that the rise of either drone technology or AI will do much to affect its basis parameters, namely the basic adequacy of existing legal principles. For the debate as such, see *inter alia* T Meron, 'Customary Humanitarian Law Today: From the Academy to the Courtroom' in A Clapham and P Gaeta (eds), *The Oxford Handbook of International Law in Armed Conflict* (2014); MN Schmitt and S Watts, 'State Opinio Juris and International Humanitarian Law Pluralism' (2015) 91 *International Law Studies* 171–215; G Best, *Humanity in Warfare: The Modern History of the International Law of Armed Conflicts* (1980); T Meron, 'Humanization of Humanitarian Law' (2000) 94 *AJIL* 239–278.

[102] See generally Y Dinstein, 'International Humanitarian Law Research Initiative: IHL in Air and Missile Warfare' (2006) www.ihlresearch.org/amw/; Y Dinstein, 'The Laws of Air, Missile and Nuclear Warfare' (1997) 27 *Isr Y B Hum Rts* 1–16.

[103] O Manea and RO Work, 'The Role of Offset Strategies in Restoring Conventional Deterrence' (2018) *Small Wars Journal* https://smallwarsjournal.com/jrnl/art/role-offset-strategies-restoring-conventional-deterrence (hereafter Manea and Work, 'The Role of Offset Strategies'); RR Tomes, 'The Cold War Offset Strategy: Assault Breaker and the Beginning of the RSTA Revolution' (*War on the Rocks*, 20 November 2014) https://warontherocks.com/2014/11/the-cold-war-offset-strategy-assault-breaker-and-the-beginning-of-the-rsta-revolution/ (hereafter Tomes, 'The Cold War Strategy').

[104] 'Since we cannot keep the United States an armed camp or a garrison state, we must make plans to use the atom bomb if we become involved in a war.' President Eisenhower in 1953, quoted in Martinage, 'Toward a New Offset Strategy' (n 96) 8. I have provided a brief history of the dynamic development of US nuclear strategy in E Afsah, 'Creed, Cabal, or Conspiracy: Origins of the Current Neo-Conservative Revolution in US Strategic Thinking' (2003) *GLJ* 902, 907–910; a fuller, accessible account can be found in DM Lawson and DB Kunsman, 'A Primer on US Strategic Nuclear Policy' (*OSTI*, 1 January 2001) www.osti.gov/servlets/purl/776355/ (hereafter Lawson and Kunsman, US Strategic Nuclear Policy').

> this transformation might occur was anyone's guess but many American observers presumed that this emerging form of warfare would proliferate rather quickly. Not widely foreseen in the mid-1990s was that nearly two decades later long-range precision strike would still be a virtual monopoly of the US military.[105]

Written in 2013, this assessment is no longer accurate. Today, a number of states have caught up and dramatically improved both the precision and range of their power projection. The gradual loss of its relative monopoly with respect to precision strike capability, remote sensing, and stealth, while simultaneously exclusive assets like aircraft carrier groups are becoming vulnerable, ineffective, or fiscally unsustainable,[106] led the US to declare its intention to respond with a 'Third Offset' strategy. It announced in 2014 that it would counter potential adversaries asymmetrically, rather than system by system:

> Trying to counter emerging threats symmetrically with active defenses or competing 'fighter for fighter' is both impractical and unaffordable over the long run. A third offset strategy, however, could offset adversarial investments in A2/AD [anti-access/area denial] capabilities in general – and ever-expanding missile inventories in particular – by leveraging US core competencies in unmanned systems and automation, extended-range and low-observable air operations, undersea warfare, and complex system engineering and integration. A GSS [global surveillance and strike] network could take advantage of the interrelationships among these areas of enduring advantage to provide a balanced, resilient, globally responsive power projection capability.[107]

The underlying developments have been apparent for some time, 'disruptive technologies and destructive weapons once solely possessed by advanced nations' have proliferated and are now easily and cheaply available to a large number of state and non-state opponents, threatening the effectiveness of many extremely expensive weapon systems on which power-projection by advanced nations, especially the US, had relied.[108] One of these disruptive technologies has been unmanned vehicles, especially airborne 'drones'. While these have been used for a century and have been militarily effective for half a century,[109] the explosion in surveillance and reconnaissance capability afforded by AI, and the dramatic miniaturisation and commercialisation of many of the underlying key components have transformed the global security landscape by making these capabilities far more accessible.[110]

Drones have proven their transformative battlefield impact since the 1973 Yom Kippur War and 1982 Israeli invasion of Lebanon.[111] Whatever their many operational and strategic benefits, unmanned aircraft were initially not cheaper to operate than conventional ones: 'higher costs for personnel needed to monitor and analyze data streams that do not exist on manned platforms, as well as the costs for hardware and software that go into the sensor packages,'[112] to say nothing of

[105] BD Watts, 'The Evolution of Precision Strike' (*CSBA*, 2013) https://csbaonline.org/uploads/documents/Evolution-of-Precision-Strike-final-v15.pdf 1–2, references omitted.
[106] Martinage, 'Toward a New Offset Strategy' (n 96) 17–20, 72.
[107] Martinage, 'Toward a New Offset Strategy' (n 96) 72.
[108] US Defence Secretary Chuck Hagel outlined these threats in a programmatic speech on 3 September 2014, which explicitly drew an analogy to Eisenhower's 'first offset' strategy and committed the country to invest in asymmetric, high-technology counter-measures, including AI, see *inter alia* Martinage, 'Toward a New Offset Strategy' (n 96) i.
[109] Their history is well summarised in Kakaes, 'From Orville Wright to September 11' (n 100).
[110] See generally Bergen and Rothenberg, *Drone Wars* (n 99).
[111] Kakaes, 'From Orville Wright to September 11' (n 100) 375.
[112] J Abizaid and R Brooks, 'Recommendations and Report of the Task Force on US Drone Policy' (*Stimson*, April 2015) www.stimson.org/wp-content/files/file-attachments/recommendations_and_report_of_the_task_force_on_us_drone_policy_second_edition.pdf 23.

the considerable expense of training their pilots,[113] left drones and the long-range precision targeting capability they conferred out of the reach of most armies, primarily due to economic costs, skilled manpower shortages, and technological complexity.

The recent conflict between Azerbaijan and Armenia has decisively shown that these conditions no longer hold. Both are relatively poor nations with fairly unsophisticated armed forces, with the crucial suppliers being the medium powers of Turkey and Israel. This highlighted the dramatic availability and affordability of such technology,[114] much of it off-the-shelf and available through a number of new entrants in the market, raising important questions of export controls and procurement.[115] Drone technology and their transformational impact on the battlefield are no longer the prerogative of rich industrial nations. While AI does not appear to have played a large role in this conflict yet,[116] the decisiveness of the precision afforded by long-range loitering munition, unmanned vehicles, and drastically better reconnaissance,[117] has not been lost on more traditional great powers.[118]

This proliferation of precision long-range weaponry portends the end of the enormous advantages enjoyed by the US as a result of its 'Second Offset' strategy. Following the Vietnam War, the US successfully sought to counteract the perceived[119] numerical superiority of the Soviet Union[120] in air and missile power by investing in superior high-precision weaponry, harnessing the country's broad technological edge.[121] These investments paid off and conferred a surprisingly long-lasting dominance. The loss of its main adversary and the inability of other adversaries to match its technological capabilities, meant that the unique advantages conferred to the US – primarily the ability to essentially eliminate risk to one's own personnel by striking remotely and to reduce political risk from 'collateral damage' by striking precisely – created an enduring willingness to deploy relatively unopposed in a vast number of unconventional conflict scenarios, sometimes dubbed a 'New American Way of War'.[122]

In principle, 'combat drones and their weapons systems are lawful weapons'.[123] Moreover, given inherent technical differences, especially their drastically higher loitering ability, lack of risk to personnel and higher precision, can actually improve observance of the law of armed

[113] Since 2009, the US Air Force has trained more drone than conventional pilots and the US Navy has announced in 2015 that the current F-35 will be the last manned strike fighter aircraft they will buy and operate, discussed in Solis, *The Law of Armed Conflict* (n 22) 547.

[114] The Turkish Bayraktar TB2 drone relies heavily on commercial civilian components, such as generic Garmin navigation systems. The UK defence minister remarked with respect to Turkey's new role as a supplier of weaponry, training, and intelligence that 'other countries are now leading the way' and that, therefore, the UK would itself begin to invest in such new, much cheaper drone technology; D Sabbagh, 'UK Wants New Drones in Wake of Azerbaijan Military Success' *The Guardian* (29 December 2020) www.theguardian.com/world/2020/dec/29/uk-defence-secretary-hails-azerbaijans-use-of-drones-in-conflict (hereafter Sabbagh, 'UK Wants New Drones').

[115] J Detsch, 'The US Army Goes to School on Nagorno-Karabakh Conflict – Off-the-Shelf Air Power Changes the Battlefield of the Future' Foreign Policy (30 March 2021) https://foreignpolicy.com/2021/03/30/army-pentagon-nagorno-karabakh-drones/.

[116] Ibid.

[117] Shaikh and Wes, 'Lessons for the Future of Strike and Defense' (n 97).

[118] Sabbagh, 'UK Wants New Drones' (n 114).

[119] There is good reason to doubt that this perceived inferiority actually existed, see Martinage, 'Toward a New Offset Strategy' (n 96) 11 *et seq*; Lawson and Kunsman, 'US Strategic Nuclear Policy' (n 104) 51–64.

[120] Manea and Work, 'The Role of Offset Strategies' (n 103).

[121] R Grant, 'The Second Offset' Air Force Magazine (24 June 2016) www.airforcemag.com/article/the-second-offset/; Tomes, 'The Cold War Offset Strategy' (n 103).

[122] RR Tomes, *US Defence Strategy from Vietnam to Operation Iraqi Freedom: Military Innovation and the New American Way of War, 1973–2003* (2006).

[123] Solis, The Law of Armed Conflict (n 22) 551.

conflict by making it easier to distinguish and reduce 'collateral damage',[124] having led some to claim that not to use drones would actually be unethical.[125] Given vastly better target reconnaissance and the possibility for much more deliberate strike decisions, convincing arguments can be made that remotely operated combat vehicles are not only perfectly lawful weapons but have the potential to increase compliance with humanitarian objectives: 'While you can make mistakes with drones, you can make bigger mistakes with big bombers, which can take out whole neighborhoods. A B-2 [manned bomber] pilot has no idea who he is hitting; a drone pilot should know exactly who he is targeting.'[126] These very characteristics – the absence of risk to military personnel and vastly better information about battlefield conditions – have also made drone warfare controversial, aspects that are heightened but not created by the addition of AI. The relative absence of operational and political risk led to a greater willingness to use armed force as a tool of statecraft, in the process bending or breaking traditional notions of international law and territorial integrity.[127] Some have argued that remote warfare with little to no risk to the operator of the weapon is somehow unethical, somehow incompatible with the warrior code of honour, concerns that should, if anything, apply even more forcefully to machines killing autonomously.[128] Whatever the merits of the conception of fairness underlying such conceptions, such 'romantic and unrealistic views of modern warfare' do not reflect a legal obligation to expose oneself to risk.[129]

There is a legal obligation, however, to adequately balance risks resulting from obtaining military advantages, which include reducing exposing service-members to risk, and the principle of distinction meant to protect innocent civilians. Many years ago, *Stanley Hoffmann* denounced the perverse doctrine of 'combatant immunity' in the context of high altitude bombing by manned aircraft staying above the range of air defences despite the obvious costs in precision and thus civilian casualties this would entail.[130] In some respects, the concerns *Hoffmann* expressed have been addressed by unmanned aircraft, which today permit unprecedented levels of precision, deliberation, and thus observance of the principle of distinction:

> Drones are superior to manned aircraft, or artillery, in several ways. Drones can gather photographic intelligence from geographic areas too dangerous for manned aircraft. Drones carry no risk of friendly personnel death or capture. Drones have an operational reach greater than that of aircraft, allowing them to project force from afar in targets far in excess of manned aircraft. The accuracy of drone-fired munitions is greater than that of most manned aircraft, and that accuracy allows them to employ munitions with a kinetic energy far less than artillery or close air support require, thus reducing collateral damage.[131]

[124] Ibid, 551–553.

[125] B Wittes, 'Drones and Democracy: A Response to Firmin DeBrabander' (*Lawfare Blog*, 15 September 2014) www.lawfareblog.com/drones-and-democracy-response-firmin-debrabander.

[126] DE Sanger, Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power (2012) 257 (hereafter Sanger, Confront and Conceal), quoted in Solis, *The Law of Armed Conflict* (n 22) 554.

[127] From the copious literature, see *inter alia* Y Dinstein, 'Concluding Observations: The Influence of the Conflict in Iraq on International Law' in RA Pedrozo (ed), *The War in Iraq: A Legal Analysis* (2010); M Sassòli, 'Ius ad Bellum and Ius in Bello: The Separation between the Legality of the Use of Force and Humanitarian Rules to be Respected in Warfare: Crucial or Outdated' in MN Schmitt and J Pejic (eds), *International Law and Armed Conflict: Exploring the Faultlines* (2007).

[128] See generally C Heyns, 'Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016); GS Corn, 'Autonomous Weapons Systems' (n 72).

[129] Solis, *The Law of Armed Conflict* (n 22) 553.

[130] S Hoffmann, 'The Politics and Ethics of Military Intervention' (1995) 37 *Survival* 29.

[131] Solis, *The Law of Armed Conflict* (n 22) 550.

At the same time, however, the complete removal of risk to one's own personnel has reduced traditional inhibitions to engage in violence abroad,[132] including controversial policies of 'targeted killings'.[133] Many of the ethical and legal *conundra*, as well as operational advantages that ensured are heightened if the capability of remotely operated vehicles is married with AI, which can improve independent or pre-authorised targeting by machines.[134]

VI. RECONNAISSANCE

The previous section showed that the rapid development of AI is transforming existing military capabilities, leading to considerable adjustments in relative strength. As in the civilian field, the main driver is the removal of a key resource constraint, namely the substitution of skilled, thus expensive and often rare, manpower by machines no longer constrained by time, availability, emotions, loyalty, alertness, etc. The area where these inherent advantages are having the largest national security impact is reconnaissance and intelligence collection.[135]

It is not always easy to distinguish these activities clearly from electronic espionage, sabotage, and intellectual property theft discussed above, but it is apparent that the capabilities conferred by automated analysis and interpretation of vast amounts of sensor data is raising important regulatory questions related to privacy, territorial integrity, and the interpretation of classical ius in bello principles on distinction, proportionality, and military necessity.

The advantages of drones outlined just above[136] have conferred unprecedented abilities to pierce the 'fog of war' by giving the entire chain of command, from platoon to commander in chief, access to information of breathtaking accuracy, granularity, and actuality.[137] Such drone-supplied information is supplemented by enormous advances in 'signal and electronic intelligence', that is eavesdropping into communication networks to obtain information relevant for tactical operations and to make strategic threat assessments. But all this available information would be meaningless without someone to make sense of it. Just like in civilian surveillance,[138] the limiting factor has long been the human being needed to watch and interpret the video or

[132] Sanger, *Confront and Conceal* (n 126).

[133] A Barak, 'International Humanitarian Law and the Israeli Supreme Court' (2014) *Isr L Rev* 181; N Melzer, *Targeted Killing in International Law* (2008); J Ulrich, 'The Gloves Were Never On: Defining the President's Authority to Order Targeted Killing in the War against Terrorism' (2005) *Va J Int'l L* 1029; D Kretzmer, 'Targeted Killings of Suspected Terrorists: Extra-Judicial Execution or Legitimate Means of Defence?' (2005) 16 *EJIL* 171.

[134] P Kalmanovitz, 'Judgment, Liability and the Risks of Riskless Warfare' in C Kreß and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (2016); Saxon, 'A Human Touch' (n 72).

[135] See also G Allen and T Chan, 'Artificial Intelligence and National Security' (*Belfer Center*, July 2017) www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf 27–35 (hereafter Allen and Chan, 'Artificial Intelligence and National Security').

[136] Solis, *The Law of Armed Conflict* (n 22) 550.

[137] See further S Smagh, 'Intelligence, Surveillance, and Reconnaissance Design for Great Power Competition' (*Congressional Research Service*, 4 June 2020) https://crsreports.congress.gov/product/pdf/R/R46389.

[138] The human factor is not only expensive and rare, it is also susceptible to bias, emotional attachment, and similar factors, which limit systemic reliability as a whole. The enormous human cost in both effort and emotional distortion in classical surveillance has been described with great artistic sensibility in the film *The Lives of Others* about the East German surveillance system. The film's great impact and merit lay in its humanisation of those charged with actually listening to the data feed; C Dueck, 'The Humanization of the Stasi in 'Das Leben der Anderen'' (2008) *German Studies Review* 599; S Schmeidl, 'The Lives of Others: Living Under East Germany's "Big Brother" or the Quest for Good Men (Das Leben der Anderen) (review)' (2009) *HRQ* 557.

data feed.[139] As this limiting factor is increasingly being removed by computing power and algorithms, real-time surveillance at hitherto impractical levels becomes possible.[140]

Whether the raw data is battlefield reconnaissance, satellite surveillance, signal intelligence, or similar sensor data, the functional challenge, regulatory difficulty, and corresponding strategic opportunity are the same: mere observation is relatively inconsequential – from both a regulatory and operational point of view – unless the information is recorded, classified, interpreted, and thereby made 'useful'.[141] This reflects a basic insight made already some forty years ago by *Herbert Simon*:

> in an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.[142]

In systems design, whether military or civilian, the main design problem is often seen as acquiring and presenting more information, following the traditional mental model that information scarcity is the chief constraint. As *Simon* and others correctly pointed out, however, these design parameters fundamentally mistake the underlying transformation brought about by technological change that is the ever-decreasing cost of collecting and transmitting data leading to the potential for 'information overload'. In other words, the real limiting factor was attention, defined as 'focused mental engagement on a particular item of information. Items come into our awareness, we attend to a particular item, and then we decide whether to act.'[143]

The true distinguishing, competitive ability is, therefore, to design systems that filter out irrelevant or unimportant information and to identify among a vast amount of data those patterns likely to require action. AI is able to automatise this difficult, taxing, and time-consuming process, by spotting patterns of activity in raw data and bringing it to the attention of humans. The key to understanding the transformation wielded by AI, especially machine learning, is the revolutionary reversal of the role of information. For most of human history, information was a scarce resource, which had to be obtained and transmitted at great material and human cost. Technological advances during the latter half of the twentieth century reversed that historic trajectory, making information suddenly over-abundant. Today, the limiting factor is no longer the availability of information as such, but our ability to make sense of its sheer amount. The ability to use computing power to sift through that sudden information abundance thus becomes a chief competitive ability, in business just as on the battlefield: 'Data mining is correctly defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.'[144] The key to performance, whether military or economic, is to derive

[139] 'The Intelligence Agencies of the United States each day collect more raw intelligence data than their entire workforce could effectively analyze in their combined lifetimes.' Allen and Chan, 'Artificial Intelligence and National Security' (n 134) 27, referring to P Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (2015) 19.

[140] This early realisation was made by Joseph Weizenbaum, the creator of ELIZA, one of the earliest natural language processing softwares. It ran on ordinary personal computers and, despite its simplicity, yielded important insights about computers themselves as social objects. The insight about surveillance was expressed in J Weizenbaum, *Computer Power and Human Reason: From Calculation to Judgment* (1976) 272.

[141] R Calo, 'Peeping HALs: Making Sense of Artificial Intelligence and Privacy' (2010) *European Journal of Legal Studies* 168, 171–174.

[142] HA Simon, *Designing Organizations for an Information-Rich World* (1971) 40–41.

[143] T Davenport and J Beck, *The Attention Economy: Understanding the New Currency of Business* (2001) 20.

[144] Zarsky, 'Mine Your Own Business!' (n 11) 4, 6.

knowledge from data, that is the ability to search for answers in complex and dynamic environments, to spot patterns of sensitive activity among often unrelated, seemingly innocuous information and to bring it to the attention of human decision-makers or initiate automated responses. Drastic advances in AI, made possible by the triple collapse in the price of sensor data collection, data storage, and processing power,[145] finally seem to offer a solution to the problem of information over-abundance by substituting machine attention for increasingly scarce human mental energy.

These long-gestating technological capabilities have suddenly aligned to bring about the maturation of AI. As we saw with respect to unmanned vehicles, one of their key structural advantages consists in their ability to deliver large amounts of sensor data, just like signal intelligence. Traditionally, one of the key constraints consisted in the highly skilled, thus rare and expensive, manpower necessary to make sense of that data: interpreting photographic intelligence, listening in on air control communications in foreign languages, etc.[146] Most of these tasks can already successfully be carried out by narrow AI, offering three game-changing advantages: first, the complete removal of manpower constraint in classifying and interpreting data, detecting patterns and predicting outcomes; second, machine intelligence is quicker than humans, it doesn't tire, it isn't biased,[147] but perhaps most importantly, it can detect patterns humans wouldn't be able to see; and third, AI permits disparate data to be fused, permitting otherwise invisible security-relevant connections to be identified.[148]

VII. FOREIGN RELATIONS

Perhaps more important than the ability to lift the 'fog of war' through better reconnaissance might be the transformation of the role of information and trust in the conduct of foreign relations. Again, this aspect of AI overlaps but is distinct from the Internet. To highlight the enormity of the challenges posed by AI, it might be useful to recall the early years of the Internet. The first time I surfed the web was in the autumn of 1995. Email was known to exist but it was not used by anyone I knew; my own first email was only sent two years later in graduate school. That autumn, I had to call and book a time-slot at the central library of the University of London, the websites I managed to find were crude, took a god-awful time to load and one had to know their addresses or look them up in a physical, printed book.[149]

My conclusion after that initial experience seemed clear: this thing would not catch on. I did not use it again for several years. After all, who would want to read a newspaper on a computer, waiting forever and scrambling through terrible layout? In a now-hilarious appearance on an American late-night show that year, the Microsoft founder *Bill Gates* responded to the host's

[145] Allen and Chan, 'Artificial Intelligence and National Security' (n 135) 14.

[146] During my graduate training at the Kennedy School of Government's specialisation in international security, my tutorial group consisted largely of seconded military officers, many of whom had been trained to do precisely these very difficult, very taxing, and fairly boring intelligence tasks. Especially the need to do this in difficult foreign languages was a very serious limiting factor. The promise of AI and especially machine learning in voice recognition etc. here is apparent.

[147] The issue of bias in the underlying algorithms is itself a field of intense scrutiny, see *inter alia* OA Osoba and W Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence* (2017).

[148] The ability of disparate, seemingly innocuous information to reveal striking and strikingly-accurate predictions has been described in a seminal newspaper article about early commercial algorithmic prediction, the principles of which have direct national security implications, see C Duhigg, 'How Companies Learn Your Secrets' *New York Times* (16 February 2012) www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[149] E Smith, 'The Internet on Dead Trees' (*Tedium*, 29 June 2017) https://tedium.co/2017/06/29/90s-internet-books-history/ (hereafter Smith, 'The Internet on Dead Trees').

thinly-disguised dismissal by giving a fairly enduring definition of that 'internet thing': 'Well, it's becoming a place where people are publishing information. … It is the big new thing.'[150] Obviously, *Gates* was more clairvoyant than me. Indeed, the Internet would be the new big thing, but he understood that it would take some time until normal people like me could see its value.[151]

Even after search-engines made the increasingly graphical web far more user-friendly, by 2000 the internet was still not mainstream and some journalists wondered whether it was 'just a passing fad'.[152] Like many new cultural phenomena driven by technological innovation, those 'in the know' enjoyed their avant-garde status, as the editor of one of the early magazines serving this new demographic stressed: 'Internet Underground was this celebration of this relatively lawless, boundless network of ideas we call the Internet. It assumed two things about its audience: 1) You were a fan [and] 2) you knew how to use it. Otherwise, the magazine wouldn't have made much sense to you.'[153] The removal of physical, temporal, and pecuniary barriers to the sharing of information indeed created a 'network of ideas', opening new vistas to collective action, new interpretations of established civil liberties, and new conceptions of geography.[154] Early generations of technophiles 'in the know' conjured this non-corporeal geography as a utopia of unfettered information-sharing, non-hierarchical self-regulation, and self-realisation through knowledge. Then-prevailing conceptions of 'cyberspace' were characterised by scepticism of both government power and commercial interests, often espousing anarchist or libertarian attitudes towards community, seeing information as a commodity for self-realisation, not profit.[155]

Early utopians stressed the opportunities created by this new, non-hierarchical 'network of ideas', which many perceived to be some kind of 'samizdat on steroids', subversive to authoritarian power and its attempts to control truth:[156] 'The design of the original Internet was biased in

[150] 'What Is Internet? Explained by Bill Gates 1995, David Letterman Show' (17 November 2019) https://youtu.be/gipL_CEw-fk, emphasis added.

[151] For the purposes of this chapter, we can ignore that he himself turned out to have misjudged how much ordinary people would see value in that Internet thing.

[152] J Chapman, 'Internet "May Be Just a Passing Fad as Millions Give Up On It"' (5 December 2000) *Daily Mail*.

[153] Rob Bernstein quoted in Smith, 'The Internet on Dead Trees' (n 149).

[154] The work of the *Electronic Frontier Foundation* illustrated the spatial metaphor and combines all three aspects that is the perceived need to defend old and necessary new rights through joint political advocacy on the frontier between traditional physical political communities and the non-corporeal space created through electronic communication, https://www.eff.org/de.

[155] S Binkley, 'The Seers of Menlo Park: The Discourse of Heroic Consumption in the 'Whole Earth Catalog'' (2003) *Journal of Consumer Culture* 283; L Dembart, '"Whole Earth Catalog" Recycled as "*Epilog*"' *New York Times* (8 November 1974) https://www.nytimes.com/1974/11/08/archives/-whole-earth-cataog-recycled-as-epilog-new-group-to-serve.html.

[156] Samizdat describes the analog distribution of unauthorised, critical literature throughout the former Communist countries using mimeographs, photocopiers, often simply re-typed carbon-copies or audio-cassettes for music or poetry readings. The effect of such underground criticism on the stability and legitimacy of the Soviet system has been devastating. Islamists used similar methods during the Iranian revolution. The advent of hard-to-monitor electronic communication portended highly destabilising times for local autocrats, but these hopes did not materialise. On the former aspect, see T Glanc, *Samizdat Past & Present* (2019); L Aron, 'Samizdat in the 21st Century' (2009) *Foreign Pol'y* 131; on the role of audio-cassettes and radio in the Iranian revolution, see BBC Persian Service, 'The History of the Revolution [انقلاب داستان]' (n.d.), www.bbc.com/persian/revolution; E Abrahamian, 'The Crowd in the Iranian Revolution' (2009) *Radical History Review* 13–38; on the role of the Internet in post-Communist politics, see S Kulikova and DD Perlmutter, 'Blogging Down the Dictator? The Kyrgyz Revolution and Samizdat Websites' (2007) *International Communication Gazette* 29–50; L Tsui, 'The Panopticon as the Antithesis of a Space of Freedom: Control and Regulation of the Internet in China' (2003) *China Information* 65; on the political space created by electronic communication generally, see JM Balkin, 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society' (2004) *NYU Law Review* 1; O Tkacheva and others, *Internet Freedom and Political Space* (2013); D Joyce, 'Internet Freedom and Human Rights' (2015) 26 *EJIL* 493.

favor of decentralization of power and freedom to act. As a result, we benefited from an explosion of decentralized entrepreneurial activity and expressive individual work, as well as extensive participatory activity. But the design characteristics that underwrote these gains also supported cybercrime, spam, and malice.'[157] Civilian internet pioneers extrapolated from these core characteristics of decentralisation and unsupervised individual agency a libertarian utopia in the true meaning of the word, a non-place or 'virtual reality' consisting of and existing entirely within a 'network of ideas'. Here, humans could express themselves freely, assume new identities and interests. Unfettered by traditional territorial regimes, new norms and social mores would govern their activities towards personal growth and non-hierarchical self-organisation. Early mainstream descriptions of the Internet compared the novelty to foreign travel, highlighting emotional, cultural, and linguistic barriers to understanding:

> The Internet is the virtual equivalent of New York and Paris. It is a wondrous place full of great art and artists, stimulating coffee houses and salons, towers of commerce, screams and whispers, romantic hideaways, dangerous alleys, great libraries, chaotic traffic, rioting students and a population that is rarely characterized as warm and friendly. . . . First-time visitors may discover that finding the way around is an ordeal, especially if they do not speak the language.[158]

As the Internet became mainstream and eventually ubiquitous, many did, in fact, learn to 'speak its language', however imperfectly.[159] The advent of AI can be expected to bring changes of similar magnitude, requiring individuals and our governing institutions to again 'learn its language'. AI is altering established notions of verification and perceptions of truth. The ability to obtain actionable intelligence despite formidable cultural and organisational obstacles,[160] is accompanied by the ability to automatically generate realistic photographs, video, and text, enabling information warfare of hitherto unprecedented scale, sophistication, and deniability.[161] Interference in the electoral and other domestic processes of competing nations are not new, but the advent of increasingly sophisticated AI is permitting 'social engineering' in novel ways.

First, it has become possible to attack large numbers of individuals with highly tailored misinformation through automated 'chatbots' and similar approaches. Secondly, the quality of 'deep fakes' generated by sophisticated AI are increasingly able to deceive even aware and skilled individuals and professional gatekeepers.[162] Thirdly, the well-known 'Eliza-effect' of human beings endowing inanimate objects like computer interfaces with human emotions, that is imbuing machines with 'social' characteristics permits the deployment of apparently responsive agents at scale, offering unprecedented opportunities and corresponding risks not only for

[157] Benkler, 'Degrees of Freedom' (n 42) 18, 19.

[158] PH Lewis, 'Personal Computers: First-Time Tourists Need a Pocket Guide to Downtown Internet' *New York Times* (5 April 1994) www.nytimes.com/1994/04/05/science/personal-computers-first-time-tourists-need-a-pocket-guide-to-downtown-internet.html; Lewis' reference to Paris and New York was probably not a coincidence, given the somewhat fearsome reputation the inhabitants of these two cities have earned, because he goes on to warn: 'Newcomers to the Internet are warned repeatedly to avoid annoying the general population with their questions.'

[159] Y Benkler, R Faris, and H Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (2018) (hereafter Benkler, Faris, and Roberts, *Network Propaganda*).

[160] B Hubbard, F Farnaz, and R Bergman, 'Iran Rattled as Israel Repeatedly Strikes Key Targets' *New York Times* (20 April 2021) www.nytimes.com/2021/04/20/world/middleeast/iran-israeli-attacks.html.

[161] Allen and Chan, 'Artificial Intelligence and National Security' (n 135) 29–34.

[162] KM Sayler and LA Harris, 'Deep Fakes and National Security' (26 August 2020) Congressional Research Service https://apps.dtic.mil/sti/pdfs/AD1117081.pdf; DK Citron and R Chesney, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) *California Law Review* 1753.

'phishing' and 'honey trap' operations,[163] but especially to circumvent an enemy government by directly targeting its population.[164]

A distinct problem fueled by similar technological advances is the ability to impersonate representatives of governments, thereby undermining trust and creating cover for competing narratives to develop.[165] Just as with any other technology, it is reasonable to expect that eventually corresponding technological advances will make it possible to detect and defuse artificially created fraudulent information.[166] It is furthermore reasonable to expect that social systems will likewise adapt and create more sophisticated consumers of such information better able to resist misinformation. Such measures had been devised during wars and ideological conflicts in the past and it is therefore correct to state that 'deep fakes don't create new problems so much as make existing problems worse'.[167] *Jessica Silbey* and *Woodrow Hartzog* are, of course, correct that the cure to the weaponisation of misinformation lies in strengthening and creating institution tasked with 'gatekeeping' and validation:

> We need to find a vaccine to the deep fake, and that will start with understanding that authentication is a social process sustained by resilient and inclusive social institutions. … it should be our choice and mandate to establish standards and institutions that are resilient to the con. Transforming our education, journalism, and elections to focus on building these standards subject to collective norms of accuracy, dignity, and democracy will be a critical first step to understanding the upside of deep fakes.[168]

The manner in which this is to be achieved goes beyond the scope of this chapter, but it is important to keep in mind that both accurate information itself, as well as misinformation have long been part of violent and ideological conflict.[169] Their transformation by the advent of AI must, therefore, be taken into account for a holistic assessment of its impact on national security and its legal regulation. This is particularly pertinent due to the rise of legal argumentation not only as a corollary of armed conflict but as its, often asymmetric, substitute in the form of 'lawfare',[170] as well as the evident importance of legal standards for such societal 'inocculation' to be successful.[171]

[163] Forsvarsministeriet, 'Center for Cybersikkerhed' (18 September 2020) https://www.fmn.dk/da/arbejdsomraader/cybersikkerhed/center-for-cybersikkerhed/.

[164] On such 'information attacks,' see generally MJ Blitz, 'Lies, Line Drawing, and (Deep) Fake News' (2018) 72 *Okla L Rev* 59; Benkler, Faris, and Roberts, Network Propaganda (n 159).

[165] S Agarwal and others, 'Protecting World Leaders against Deep Fakes' (2019) *IEEE Xplore* 38.

[166] For an account of the technology involved, see for instance S Agarwal and others, 'Detecting Deep-Fake Videos from Appearance and Behavior' (2020) *IEEE International* 1.

[167] J Silbey and W Hartzog, 'The Upside of Deep Fakes' (2019) 78 *Maryland Law Review* 960, 960.

[168] Ibid, 966.

[169] R Darnton, 'The True History of Fake News' *The New York Review* (13 February 2017) www.nybooks.com/daily/2017/02/13/the-true-history-of-fake-news/.

[170] The term has been suggested by General Charles Dunlap who offered the following definition: 'the strategy of using – or misusing – law as a substitute for traditional military means to achieve a warfighting objective.' CJ Dunlap, 'Lawfare Today: A Perspective' (2008) *Yale J Int'l L* 146. 146. See also D Stephens, 'The Age of Lawfare', in RA Pedrozo and DP Wollschlaeger (eds), *International Law and the Changing Character of War* (2011); CJ Dunlap, 'Lawfare Today … and Tomorrow', in RA Pedrozo and DP Wollschlaeger (eds), *International Law and the Changing Character of War* (2011).

[171] See *inter alia* Chapter 13 "What Can Men Do against Such Reckless Hate?" in Benkler, Faris, and Roberts, *Network Propaganda* (n 159) 351–380.

VIII. ECONOMICS

National security is affected by economic competitiveness, which supplies the fiscal and material needs of military defence. The impact of the ongoing revolution in AI on existing labour markets and productive patterns is likely to be transformational.[172] The current debate is reminiscent of earlier debates about the advent of robotics and automation in production. Where that earlier debate focused on the impact on the bargaining power and medium-term earning potential of blue-collar workers, AI is also threatening white-collar workers, who hitherto seemed relatively secure from cross-border wage arbitrage as well as automation.[173] In a competitive arena, whether capitalism for individual firms or anarchy for nations, the spread of innovation is not optional but a logical consequence of the 'socialising effect' of any competitive system:[174] 'Machine learning is a cool new technology, but that's not why businesses embrace it. They embrace it because they have no choice.'[175]

This embrace of AI has at least three important national security implications, with corresponding regulatory challenges and opportunities. First, dislocations resulting from the substitution of machines for human labour has destabilising effects for social cohesion and political stability, both domestic and international.[176] These dislocations have to be managed, including through the use of proactive regulation meant to further positive effects while buffering negative consequences.[177] The implications of mass unemployment resulting from this new wave of automation is potentially different from earlier cycles of technological disruption because it could lead to permanent unemployability of large sectors of the population, rendering them uncompetitive at any price. This could spell a form of automation-induced 'resource curse' affecting technologically advanced economies,[178] suddenly suffering from the socio-economic-regulatory failings historically associated with underdeveloped extractive economies.[179]

Second, the mastery of AI has been identified by all major economic powers as central to maintaining their relative competitive posture.[180] Consequently, the protection of intellectual property, the creation of a conducive regulatory, scientific, and investment climate to nurture the sector has itself increasingly become a key area of competition between nations and trading blocs.[181]

[172] European Commission, 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)' (*European Commission*, 26 April 2021) https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[173] K Roose, 'The Robots Are Coming for Phil in Accounting' *New York Times* (6 March 2021) www.nytimes.com/2021/03/06/business/the-robots-are-coming-for-phil-in-accounting.html.

[174] KN Waltz, *Theory of International Politics* (1979) 129.

[175] P Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (2015) 13.

[176] Allen and Chan, 'Artificial Intelligence and National Security' (n 134) 36–39.

[177] Denmark and the other Scandinavian economies have a long history of seeking productivity gains in both the public and private sector as a way to keep their costly welfare systems fiscally sustainable and labour markets globally competitive. See *inter alia* Forsvarsministeriet, 'National strategi for cyber- og informationssikkerhed. Øget professionalisering og mere viden' (December 2014); C Greve and N Ejersbo, *Moderniseringen af den offentlige sektor* (3rd ed. 2014); J Hoff, *Danmark som Informationssamfund. Muligheder og Barrierer for Politik og Demokrati* (2004); PA Hall, 'Danish Capitalism in Comparative Perspective', in JL Campbell, JA Hall, and OK Pedersen (eds), *National Identity and the Varieties of Capitalism: The Danish Experience* (2006).

[178] Allen and Chan, 'Artificial Intelligence and National Security' (n 135) 37.

[179] See *inter alia* G Luciani, 'Allocation v Production States: A Theoretical Framework' in G Luciani and B Hazem (eds), *The Rentier State* (2015).

[180] Mozur and Myers, 'Xi's Gambit' (n 5); R Doshi and others, 'China as a "Cyber Great Power" – Beijing's Two Voices in Telecommunications' (*Brookings*, April 2021) www.brookings.edu/wp-content/uploads/2021/04/FP_20210405_china_cyber_power.pdf.

[181] Bird and others, 'The Ethics of Artificial Intelligence' (n 59).

Third, given the large overlap between civilian and military sectors, capabilities in AI developed in one are likely to affect the nation's position in the other.[182] Given inherent technological characteristics, especially scalability and the drastic reduction of marginal costs, and the highly disruptive effect AI can have on traditional military capabilities, the technology has the potential to drastically affect the relative military standing of nations quite independent of conventional measures such as size, population, hardware, etc.: 'Small countries that develop a significant edge in AI technology will punch far above their weight.'[183]

IX. CONCLUSION

Like many previous innovations, the transformational potential of AI has long been 'hyped' by members of the epistemic communities directly involved in its technical development. There is a tendency among such early pioneers to overstate potential, minimise risk, and alienate those not 'in the know' by elitist attitudes, incomprehensible jargon, and unrealistic postulations. As the comparison with cyberspace has shown, it is difficult to predict with accuracy what the likely impact of AI will be. Whatever its concrete form, AI is almost certain to transform many aspects of our lives, including national security.

This transformation will affect existing relative balances of power and modes of fighting and thereby call into question the existing normative *acquis*, especially regarding international humanitarian law. Given the enormous potential benefits and the highly dynamic current stage of technological innovation and intense national competition, the prospects for international regulation, let alone outright bans are slim. This might appear to be more consequential than it is, because much of the transformation will occur in operational, tactical, and strategic areas that can be subsumed under an existing normative framework that is sufficiently adaptable and broadly adequate.

The risk of existential danger by the emergence of super-intelligence is real but perhaps overdrawn. It should not detract from the laborious task of applying existing international and constitutional principles to the concrete regulation of more mundane narrow AI in the national security field.

[182] Allen and Chan, 'Artificial Intelligence and National Security' (n 135) 35–41.
[183] Ibid, 3 and 58–59.

27

Morally Repugnant Weaponry?

Ethical Responses to the Prospect of Autonomous Weapons

Alex Leveringhaus

I. INTRODUCTION

In 2019, the United Nations (UN) Secretary General *Antonio Guterres* labelled lethal autonomous weapons 'as political unacceptable and morally repulsive'.[1] 'Machines', *Guterres* opined, 'with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law'.[2] The Secretary General's statement seems problematic. Just because something is morally repugnant does not entail that it should be banned by law. Further, it is not clear what exactly renders autonomous weapons systems (AWS hereinafter) morally abhorrent.[3] The great danger is that statements such as the Secretary General's merely rely on the supposed 'Yuck' factor of AWS.[4] But Yuck factors are notoriously unreliable guides to ethics. While individuals might find things 'yucky' that are morally unproblematic, they might not be repulsed by things that pose genuine moral problems.

In response to the Secretary General's statement, the purpose of this chapter is twofold. First, it seeks to critically survey different ethical arguments against AWS. Because it is beyond the scope of this chapter to survey every ethical argument in this context, it outlines three prominent ones, (1) that AWS create so-called responsibility gaps; (2) that the use of lethal force by an AWS

[1] N Werkhauser, 'UN Impasse Could Mean Killer Robots Escape Regulation' DW News (20 August 2018) www.dw.com/en/un-impasse-could-mean-killer-robots-escape-regulation/a-50103038 (hereafter Werkhauser, 'Killer Robots').

[2] Secretary-General, *Machines Capable of Taking Lives without Human Involvement are Unacceptable, Secretary-General Tells Experts on Autonomous Weapons Systems* (United Nations Press Briefing, 25 March 2019), www.un.org/press/en/2019/sgsm19512.doc.htm.

[3] To avoid any misunderstanding at the outset, autonomy, in the debate on AWS, is not understood in the same way as in moral philosophy. Autonomy, in a moral sense, means to act for one's own reasons. This is clearly not the case in the context of AWS. These systems, as I point out shortly, require programming by a human individual. In quasi-Kantian parlance, then, AWS are heteronomous, rather than autonomous, in that they do not act for their own reasons. As I shall explain later, in the context of the debate on AWS, autonomy essentially describes a machine's capacity, once it has been programmed, to carry out tasks independently of, and without further guidance from, a human individual. This is, of course, not sufficient for moral autonomy in a meaningful sense. In the chapter, I use the term autonomy according to its technological meaning, rather than its moral one.

[4] The term Yuck factor describes a strong emotional reaction of revulsion and disgust towards certain activities, things, or states of affairs. The question is whether such visceral emotional responses are a reliable guide to ethics. Some activities or things – for example, in vitro meat or a human ear grown on a mouse for transplantation – might seem disgusting to some people, and sometimes this can indeed have normative significance. That being said, the feeling of disgust does not always explain why something is ethically undesirable. One problem is that our emotional responses are often shaped by social, economic, and political factors that can cloud our ethical judgement. Especially in the context of emerging technologies, the danger is that the Yuck factor might prevent the adoption of technologies that might be genuinely beneficial.

is incompatible with human dignity; and (3) that AWS replace human agency with artificial agency. The chapter contends that neither of these arguments is sufficient to show that AWS are morally repugnant. Second, drawing upon a more realistic interpretation of the technological capacities of AWS, the chapter outlines three alternative arguments as to why AWS are morally problematic, as opposed to morally repugnant.

In the second part of the chapter, I write more about definitional issues in the debate on AWS. In the third part, I critically analyse, respectively, the notion of a responsibility gap, the relationship between AWS and human dignity, and role of human agency in war. In the fourth part, I outline a brief alternative account of why AWS might be morally problematic and explain how this intersects with other key issues in contemporary armed conflict.

Before I do so, I need to raise three general points. First, the chapter does not discuss the legal status of AWS. The focus of this chapter is on ethical issues only. The question whether, as suggested by the Secretary General, the alleged moral repugnancy of AWS justifies their legal prohibition is best left for a different occasion. Second, the chapter approaches AWS from the perspective of contemporary just war theory as it has developed since the publication of *Michael Walzer's* seminal "Just and Unjust Wars: A Moral Argument with Historical Illustrations" in 1977.[5] Central to *Walzer's* work, and much of just war theory after it, is the distinction between the normative frameworks of *jus ad bellum* (justice in the declaration of war) and *jus in bello* (justice in the conduct of war). As we shall see, the ethical debate on AWS has mainly been concerned with the latter, as it has tended to focus on the use of (lethal) force by AWS during armed conflict. Third, in addition to the distinction between *jus ad bellum* and *jus in bello*, *Walzer*, in *Just and Unjust Wars*, defends the distinction between combatants (who may be intentionally killed) and non-combatants (who may not be intentionally killed) during armed conflict. The former are typically soldiers, whereas the latter tend to be civilians, though he acknowledges the existence of grey zones between these categories.[6] In recent years, this distinction has come increasingly under pressure, with some theorists seeking to replace it with a different one.[7] For the sake of convenience and because these terms are widely recognised, the chapter follows *Walzer* in distinguishing between combatants and non-combatants. However, many of the issues highlighted in the following sections will also arise for theories that are critical of *Walzer's* distinction.

II. WHAT IS AN AUTONOMOUS WEAPON?

Here, I offer a fourfold attempt to define AWS. First, it is self-evident that AWS are weapons. In this sense, they differ from other forms of (military) technology that are not classifiable as weapons. The following analysis assumes that weapons have the following characteristics; (1) they were specifically designed in order to (2) inflict harm on another party.[8] Usually, the harm is achieved via a weapon's kinetic effect. The harmful kinetic effect is not random or merely a by-product of the weapon's operation. Rather, weapons have been intentionally designed to produce a harmful effect. Non-weapons can be used as weapons – you could stab me with the butterknife – but they have not been deliberately designed to inflict harm.

[5] M Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations* (5th ed. 2015) (hereafter Walzer, *Just and Unjust Wars*).
[6] Ibid, 145.
[7] See J McMahan, *Killing in War* (2009).
[8] J Forge, *Designed to Kill: The Case against Weapons Research* (2013).

Second, as stated by Secretary General *Guterres*, the crucial feature of AWS, accounting for their alleged moral repugnancy, is that their kinetic and potentially lethal effect is created by the weapon without human involvement.[9] However, AWS will require initial mission programming by a human programmer. Hence, there will be human involvement in the deployment of an AWS. The point, though, is that once an AWS has been programmed with its mission parameters, the weapon is capable of operating without any further guidance and supervision by a human individual. Crucially, it can create a harmful and potentially lethal kinetic effect by delivering a payload without direct or real-time human involvement. The technical term for such a weapon is an out-of-the-loop system. Unlike in-the-loop-systems in which the decision to apply kinetic force to a target is made by the weapon's operator in real-time, or on-the-loop systems where the operator remains on stand-by and can override the weapon, a genuine out-of-the-loop system will not involve an operator once deployed.[10]

Third, the notion of out-of-the-loop systems could be equally applied to automated and autonomous systems. Indeed, the literature is far from clear where the difference between the two lies, and any boundaries between automated and autonomous machine behaviour might be fluid. As a rule of thumb, autonomous systems are more flexible in their response to their operating environment than automated ones.[11] They could learn from their prior experiences in order to optimise their (future) performance, for example. They might also have greater leeway in translating the orders given via their programming into action. What this means in practice is that, compared to an automated system, any autonomous system (and not just weapons) is less predictable in its behaviour. That said, AWS would be constrained by particular targeting categories. That is, their programming would only allow them to attack targets that fall within a particular category. To illustrate the point, an AWS programmed to search and destroy enemy tanks would be restricted to attacking entities that fall into this category. Yet, compared to an automated weapon, it would be hard to predict where, when, and which enemy tank it would attack.

Fourth, as the quote from Secretary General *Guterres* suggests, AWS can produce a lethal kinetic effect without any human intervention post-programming. Here, the question is whether the alleged moral repugnancy of AWS only refers to AWS that would be deliberately programmed to attack human individuals. If so, this would potentially leave scope for the development and deployment of AWS that are not used for this purpose, such as the one mentioned in the 'enemy tank' example above. Moreover, it is noteworthy that any weapon can kill in two ways, (1) as an intended effect of its operation, and (2) as a side-effect of its operation. Presumably, the earlier quote by Secretary *Guterres* refers to (1), where a programmer would intentionally programme an AWS in order to attack human individuals, most likely enemy combatants.

The focus on this issue is problematic, for two reasons. First, it neglects lethal harm that might arise as a side effect of the operation of an AWS. As I shall show later, this category of harm is, in the context of AWS, more morally problematic than intended harm. Second, it is doubtful whether the intentional targeting of individuals through AWS is legally and morally permissible. To explain, as was noted in the introduction to this chapter, at the level of *jus in bello*, contemporary just war theory post-*Walzer* rests on the distinction between combatants and non-combatants. True, given advances in machine vision, an AWS could, with great reliability,

[9] Werkhauser, 'Killer Robots' (n 1)

[10] P Scharre, *Army of None: Autonomous Weapons and the Future of War* (2019).

[11] A Leveringhaus, *Ethics and Autonomous Weapons* (2006) 46 *et seq* (hereafter Leveringhaus, *Ethics and Autonomous Weapons*).

distinguish between human individuals and non-human objects and entities. Yet, what it cannot do, at the present state of technological development at least, is to accurately determine whether an individual is a legitimate target (a combatant) or an illegitimate target (a non-combatant). It is, in fact, hard to see how a machine's capacity for such a qualitative judgement could ever be technologically achieved. As a result, the deployment of an AWS to deliberately kill human individuals would not be permissible under *jus in bello*.

If the above observation is true, it has two immediate repercussions for the debate on AWS. First, militaries might not be particularly interested in developing systems whose purpose is the autonomous targeting of human individuals, knowing that such systems would fall foul of *jus in bello*. Still, militaries may seek to develop AWS that can be programmed to attack more easily identifiable targets – for example, a tank, a missile, or a submarine. In this case, I contend that the ethical debate on AWS misses much of the actual technological development and restricts its own scope unnecessarily. Second, as I have argued elsewhere,[12] in order to assess whether programming an AWS to kill human individuals is morally repugnant, it is necessary to assume that AWS do not fall down at the normative hurdle of accurately identifying human individuals as legitimate or illegitimate targets. This assumption is a necessary philosophical abstraction and technological idealisation of AWS that may not reflect their actual development and potential uses. Bearing this in mind, the chapter continues by analysing whether it is morally repugnant to deliberately programme an AWS to kill human individuals in war.

III. PROGRAMMED TO KILL: THREE ETHICAL RESPONSES

The main ethical argument in favour of AWS is essentially humanitarian in nature.[13] More precisely, the claim is that AWS (1) ensure stricter compliance with *jus in bello*, and (2) reduce human suffering and casualties as a result.[14] Interestingly, the ethical counterarguments do not engage with this humanitarian claim directly. Rather, they immediately attack the notion of autonomous uses of force via an AWS. In this part of the chapter, I look at three ethical responses to the prospect of AWS being intentionally programmed to take human lives, (1) the argument that AWS create so-called responsibility gaps, (2) the claim that the intentional use of AWS to kill is incompatible with human dignity, and (3) the argument (made by this author) that, by replacing human agency with artificial agency at the point of force delivery, AWS render humans incapable of revising a decision to kill. As indicated above, the three arguments rely on a technologically-idealised view of AWS.

1. *Responsibility Gaps*

One of the earliest contributions to the ethical debate on AWS is the argument that these weapons undermine a commitment to responsibility. Put simply, the claim is that, in certain cases, it is not possible to assign (moral) responsibility to a human individual for an event caused by an AWS. This is especially problematic if the event constitutes a violation of *jus in bello*. In such cases, neither the manufacturer of the AWS, nor its programmer, nor the AWS itself (of course) can be held responsible for the event, resulting in a responsibility gap.[15] This gap

[12] Leveringhaus, *Ethics and Autonomous Weapons* (n 11).
[13] Ibid, 62–63.
[14] R Arkin, 'The Case for Ethical Autonomy in Unmanned Systems' (2010) 9(4) *Journal of Military Ethics*, 332–341.
[15] R Sparrow, 'Killer Robots' (2007) 24(1) *Journal of Applied Philosophy*, 62–77.

arises from the inherent unpredictability of autonomous machine behaviour. No human programmer, it is claimed, could foresee every facet of emerging machine behaviour. Hence, it is inappropriate, the argument goes, to hold the programmer – let alone the manufacturer – responsible for an unforeseen event caused by an AWS. In a moral sense, no one can be praised or blamed, or even punished, for the event. Why should this pose a moral problem? Here, the claim is that for killing in war to be morally permissible, someone needs to be held responsible for the use of force. Responsibility gaps, thus, undermine the moral justification for killing in war.

Admittedly, the idea of a responsibility gap is powerful. But it can be debunked relatively easily. First, moral responsibility can be backward-looking and forward-looking. The responsibility gap arises from a backward-looking understanding of responsibility, where it is impossible to hold a human agent responsible for an event caused by an AWS in the past. The argument has nothing to say about the forward-looking sense of responsibility, where an agent would be assigned responsibility for supervising, controlling, or caring for someone or something in the future. In the present context, the forward-looking sense of responsibility lends itself to an on-the-loop system, rather than an out-of-the-loop system. Either way, it is not clear whether a gap in backward-looking responsibility is sufficient for the existence of a responsibility gap, or whether there also needs to be a gap in forward-looking responsibility. A backward-looking gap may be a necessary condition here, but not a sufficient one.

Second, it is contested whether killing in war is *prima facie* permissible if, and only if, someone can be held responsible for the use of lethal force. There are, roughly, two traditions in contemporary moral philosophy for thinking about the issue.[16] The first, derived from Thomism, is agent-centric in that it focuses on the intentions of the agent using lethal force. The second tradition is target-centric in that it focuses on the moral status of the target of lethal force. That is to say, the permissibility centres on the question whether the target has become liable to attack because it is morally and/or causally responsible for a (unjust) threat. On the target-centric approach, an agent who could not be held responsible for the use of lethal force may be allowed to kill if the target was liable to attack. In short, then, if the link between (agent) responsibility and the moral permission to use force is far weaker than assumed, the idea of a responsibility gap loses its normative force.

Third, the idea of a responsibility gap lets those who deployed an AWS off the hook far too easily.[17] True, given that autonomous systems tend to show unpredictable emerging behaviours, the individual (or group of individuals) who deploys an AWS by programming it with its mission parameters cannot know in advance that, at t_5, the AWS is going to do x. Still, the programmer and those in the chain of command above him know that the AWS they deploy is likely to exhibit unforeseen behaviour, which might, in the most extreme circumstances, result in the misapplication of force. Notwithstanding that risk, they choose to deploy the weapon. In doing so, they impose a significant risk on those who might come into contact with the AWS in its area of operation, not least non-combatants. Of course, the imposition of that risk may either be reasonable and permissible under the circumstances or unreasonable and reckless – more on this shortly. But generally, the claim that those deploying an AWS are not responsible for any unforeseen damage resulting from its operation appears counterintuitive.

Finally, even if it is hard to hold individuals responsible for the deployment of an AWS, it is worthwhile remembering that armed conflicts are (usually) fought by states. In the end, the buck stops there. Needless to say, this raises all sorts of difficult issues which the chapter cannot go

[16] S Uniacke, *Permissible Killing: The Self-Defence Justification of Homicide* (1994).
[17] Leveringhaus, *Ethics and Autonomous Weapons* (n 11) 76–86.

into. For now, it suffices to note that states have made reparations for the (wrongful) damage they caused in armed conflict. Most recently, for instance, the United States (US) compensated Afghan civilians for the deaths of (civilian) family members in the course of US military operations in the country as part of the so-called War on Terror.[18] The most notorious case is that of Staff Sergeant *Robert Bales* who, after leaving his base without authorisation, went on a shooting rampage and was later charged with the murder of seventeen Afghan civilians, as well as causing injury to a number of others. The US paid compensation to those affected by Sergeant *Bales*' actions, even though Sergeant *Bales* acted out of his own volition and outside the chain of command.[19]

In sum, the notion of a responsibility gap does not prove that AWS are morally repugnant. Either the existence of a (backward-looking) responsibility gap is insufficient to show that the deployment of AWS would be morally unjustifiable or there is no responsibility gap as such. Yet, there are elements of the responsibility gap that could be salvaged. The argument that it is necessary to be able to hold someone responsible for the use of force is motivated by a concern for human dignity or respect for individuals. It might, therefore, be useful to focus on the relationship between AWS and human dignity. That is the purpose of the next section.

2. *Dignity*

Are AWS morally repugnant because, as has been suggested by some contributors to the debate, they are an affront to human dignity?[20] This question is difficult to answer because just war theorists have tended to eschew the concept of human dignity. Perhaps for good reason. Appeals to dignity often do not seem to decisively resolve difficult moral issues. For instance, the case for, as well as against, physician-assisted suicide could be made with reference to the concept of dignity. That said, the concept enters into contemporary just war thinking, albeit in an indirect way. This has to do with the aforementioned distinction between combatants and non-combatants. The former group is seen as a legitimate target in armed conflict, which means that combatants lack a moral claim against other belligerent parties not to intentionally kill them. Non-combatants, by contrast, are immune to intentional attack, which means that they hold a negative moral claim against combatants not to intentionally kill them. However, *jus in bello* does not grant non-combatants immunity against harm that would be unintentionally inflicted. Here, the Doctrine of Double Effect and its conceptual and normative distinction between intended and foreseen harm comes into play. In his classic discussion of non-combatant immunity, *Walzer* argues that it is permissible to kill or harm non-combatants if, and only if, the harm inflicted on them is (1) not intended, (2) merely foreseen (by the belligerent), (3) not used as a (bad) means to a good effect, (4) proportionate (not excessive to the good achieved), and (5) consistent with a belligerent's obligations of 'due care'.[21]

Granted, but why should the distinction between intended and foreseen harm have any normative significance? According to the *Kantian* view, the Doctrine of Double Effect protects

[18] See M Gluck, 'Examination of US Military Payments to Civilians Harmed during Conflict in Afghanistan and Iraq' (*Lawfare*, 8 October 2020) www.lawfareblog.com/examination-us-military-payments-civilians-harmed-during-conflict-afghanistan-and-iraq.

[19] Associated Press, 'US Compensation for Afghanistan Shooting Spree' (*The Guardian*, 25 March 2012) www.theguardian.com/world/2012/mar/25/us-compensation-afghanistan-shooting-spree.

[20] See A Pop, 'Autonomous Weapon Systems: A Threat to Human Dignity?' (International Committee of the Red Cross, *Humanitarian Law & Policy*, 10 April 2018) https://blogs.icrc.org/law-and-policy/2018/04/10/autonomous-weapon-systems-a-threat-to-human-dignity/.

[21] Walzer, *Just and Unjust Wars* (n 5) 153–154.

the dignity of innocent individuals by ensuring that belligerents comply with the second formulation of *Kant's* categorical imperative, which obliges them to treat (innocent) individuals not merely as means to an end but always also as ends-in-themselves.[22] To illustrate the point, if Tim intentionally bombs non-combatants in order to scare the enemy into surrender, Tim violates their status as ends-in-themselves, instrumentalising their deaths in order to achieve a particular goal (the end of the war). By contrast, if Tom bombs a munitions factory and unintentionally kills non-combatants located in its vicinity as a foreseen side-effect of his otherwise permissible (and proportionate) military act, Tom does not instrumentalise their deaths for his purposes. Counterfactually, Tom could destroy the munitions factory, even if no non-combatant was harmed. Unlike Tim, Tom does not need to kill non-combatants to achieve his goals. Tom's actions would not violate the ends-not-means principle – or so one might argue.

According to the *Kantian* View of the Doctrine of Double Effect, then, if Tam intentionally programmed an AWS to kill non-combatants, he would violate their dignity. Note, though, that there is no moral difference between Tam's and Tim's actions. The only difference is the means they use to kill non-combatants. As a result, this example does not show that AWS pose a unique threat to human dignity. Any weapon could be abused in the way Tam abuses the AWS. Hence, in the example, the use of the AWS is morally repugnant, not the weapon as such.

What about combatants? If Tam intentionally programmed an AWS to kill enemy combatants, would he violate their dignity? That question is hard to answer conclusively. First, because combatants lack a moral claim not to be killed, Tam does not violate their moral rights by deploying an AWS against them. Second, unlike non-combatants, it is usually morally permissible and necessary to instrumentalise combatants. One does not need to go quite as far as Napoleon who remarked that 'soldiers are made to be killed'.[23] But *Walzer* is right when he observes that combatants are the human instruments of the state.[24] As a result, combatants enjoy far lower levels of protection against instrumentalization than non-combatants. In a nutshell, it needs to be shown that, although combatants (1) lack a moral claim not to be intentionally attacked [during combat], and (2) do not enjoy the same level of protection against instrumentalization as non-combatants, the use of an AWS in order to kill them would violate their dignity.

The dignity of combatants, critics of AWS may argue, is violated because a machine should not be left to decide who lives or dies. At the macro-level of programming the argument is certainly wrong. Tam, the programmer in the above example, makes the decision to programme an AWS to detect and eliminate enemy combatants. In this sense, the machine Tam deploys does not make a decision to take life. Tam does. At the micro-level of actual operations, though, the argument has some validity. Here, the machine has some leeway in translating Tam's instructions into actions. Within the target category of enemy combatants, it could 'decide' to attack Combatant_1 rather than Combatant_2 or Combatant_3. It might, further, not be possible to ascertain why the machine chose to attack Combatant_1 over Combatant_2 and Combatant_3. The resulting question is whether the machine's micro-choice, rather than Tam's macro-choice, violates Combatant_1's dignity.

Arguably not. This is because killing in war tends to be impersonal and to some extent morally arbitrary. Why did a particular combatant die? Often, the answer will be that he was a combatant. Armed conflict, as *Walzer* observes, is not a personal relationship. To wit,

[22] TA Cavanaugh, *Double Effect Reasoning: Doing Good and Avoiding Evil* (2006).
[23] Walzer, *Just and Unjust Wars* (n 5) 136.
[24] Ibid, 36–45.

combatants are not enemies in a personal sense, which would explain the choices they make. They are the human instruments of the state. They kill and die because they are combatants. And often because they are in the wrong place at the wrong time. That is the brutal reality of warfare. Consider a case where an artillery operator fires a mortar shell in the direction of enemy positions. Any or no enemy combatant located in the vicinity might die as a result. We might never know why a particular enemy combatant died. We only know that the artillery operator carried out his orders to fire the mortar shell. By analogy, the reason for an AWS's micro-choice to target Combatant_1 over Combatant_2 and Combatant_3 is, ultimately, that Combatant_1 is a combatant. Combatant_1 was simply in the wrong place at the wrong time. It is not clear why this micro-choice should be morally different from the artillery operator's decision to fire the mortar shell. Just as the dignity of those combatants who were unlucky enough to be killed by the artillery operator's mortar shell is not violated by the artillery operator's actions, Combatant_1's dignity is not violated because a machine carried out its pre-programmed orders by micro-choosing him over another combatant. So, the argument that human dignity is violated if a machine makes a micro-choice over life and death seems morally dubious.

But perhaps critics of AWS may concede that the micro-choice as such is not the problem. To be sure, killing in war, even under orders, is to some extent random. The issue, they could reply, is that the artillery operator and those whom he targets have equal skin in the game, while the AWS that kills Combatant_1 does not. In other words, the artillery operator has an appreciation of the value of (his own) life, which a machine clearly lacks. He is aware of the deadly effects of his actions, whereas a machine is clearly not. Perhaps this explains the indignity of being killed as a result of a machine's micro-choice.

This argument takes us back to the *Thomistic* or agent-centric tradition in the ethics of killing outlined previously. Here, the internal states of the agent using force, rather than the moral status of the target, determines the permissibility of killing. To be allowed to kill in war, a combatant needs to have an appreciation of the value of life or at least be in a similar situation to those whom he targets. Naturally, if one rejects an agent-centric approach to the ethics of killing, this argument does not hold much sway.

More generally, it is unclear whether such a demanding condition – that an individual recognises the value of life – could be met in contemporary armed conflict. Consider the case of high altitude bombing during NATO's war in Kosovo. At the time, *Michael Ignatieff* observed that NATO was fighting a 'virtual war' in which NATO did the fighting while most of the Serbs 'did the dying'.[25] It is hard to imagine that NATO's bomber pilots, flying at 15,000 ft and never seeing their targets, would have had the value of human life at the forefront of their minds, or would have even thought of themselves as being in the same boat as those they targeted. The pilots received certain target coordinates, released their payloads once they had reached their destination, and then returned to their base. In short, modern combat technology, in many cases, has allowed combatants to distance themselves from active theatres, as well as the effects of their actions, to an almost unprecedented degree. These considerations show that the inability of a machine to appreciate the value of life does not pose a distinctive threat to human dignity. The reality of warfare has already moved on.

But there may be one last argument available to those who seek to invoke human dignity against AWS. To be sure, combatants, they could concede, do not hold a moral claim against other belligerents not to attack them. Nor, as instruments of the state, do they enjoy the same level of protection against instrumentalization as non-combatants. Still, unless one adopts

[25] M Ignatieff, *Virtual War* (2000).

Napoleonic cynicism, there must be some moral limits on what may permissibly be done to combatants on the battlefield. There must be some appreciation that human life matters, and that humans are not merely a resource that can be disposed of in whatever way necessary. Otherwise, why would certain weapons be banned under international law, such as blinding lasers, as well as chemical and biological weapons?

Part of the answer is that these weapons are likely to have an indiscriminate and disproportionate effect on non-combatants. But intuitively, as the case of blinding lasers illustrates, there is a sense that combatants deserve some protection. Are there certain ways of killing that are somehow cruel and excessive, even if they were aimed at legitimate human targets? And if that is the case, would AWS fall into this category?

There is a comparative and a non-comparative element to these questions. Regarding the comparative element, as macabre as it sounds, it would certainly be excessive to burn a combatant to death with a flamethrower if a simple shot with a gun would eliminate the threat he poses. That is common-sense. With regard to the non-comparative element, the issue is whether there are ways of killing which are intrinsically wrong, regardless of how they compare to alternative means of killing. That question is harder to answer. Perhaps it is intrinsically wrong to use a biological weapon in order to kill someone with a virus. That said, it is hard to entirely avoid comparative judgements. Given the damage that even legitimate weapons can do; it is not clear that their effects are always morally more desirable than those of illegitimate weapons. One wonders if it is really less 'cruel' for someone to bleed to death after being shot or to have a leg blown off from an explosive than to be poisoned. Armed conflict is brutal and modern weapons technology is shockingly effective, notwithstanding the moral (and legal) limits placed on both.

Although, within the scope of this chapter, it is impossible to resolve the issues arising from the non-comparative element, the above discussion provides two main insights for the debate on AWS. First, if AWS are equipped with payloads whose effects were either comparatively or non-comparatively excessive or cruel, they would certainly violate relevant moral prohibitions against causing excessive harm. For example, an autonomous robot with a flamethrower that would incinerate its targets or an autonomous aerial vehicle that would spray target areas with a banned chemical substance would indeed be morally repugnant. Second, it is hard to gauge whether the autonomous delivery of a legitimate – that is, not disproportionately harmful – payload constitutes a cruel or excessive form of killing. Here, it seems that the analysis is increasingly going in circles. For, as I argued above, many accepted forms of killing in war can be seen analogous to, or even morally on a par with, autonomous killing. Either all of these forms of killing are a threat to dignity, which would lend succour to ethical arguments for pacifism, or none are.

To sum up, AWS pose a threat to human dignity if they were deliberately used to kill non-combatants, or were equipped with payloads that caused excessive or otherwise cruel harm. However, even in such cases, AWS would not pose a distinctive threat. This is because some of the features of autonomous killing can also be found in established forms of killing. The moral issues AWS raise with regard to dignity are not unprecedented. In fact, the debate on AWS might provide a useful lens through which to scrutinise established forms of killing in war.

3. *Human and Artificial Agency*

If the earlier arguments are correct, the lack of direct human involvement in the operation of an AWS, once programmed, is not a unique threat to human dignity. Yet, intuitively, there is something morally significant about letting AWS kill without direct human supervision.

This author has sought to capture this intuition via the Argument from Human Agency.[26] I argue that AWS have artificial agency because they interact with their operating environment, causing changes within it. According to the Argument from Human Agency, the difference between human and artificial agency is as follows. Human agency consists in refusing to carry out an order. As history shows, soldiers have often not engaged the enemy, even when under orders to do so. An AWS, by contrast, will kill once it has 'micro-chosen' a human target. We might not know when, where, and whom it will kill, but it will carry out its programming. In a nutshell, by removing human agents from the point of payload delivery, out-of-the-loop systems make it impossible to revise a decision to kill.

While the Argument from Human Agency captures intuitions about autonomous forms of killing, it faces three challenges. First, as was observed above, combatants do not hold a moral claim not to be killed against other belligerent parties and enjoy lower levels of protection against instrumentalization than non-combatants. Why, then, critics of the Argument from Human Agency might wonder, should combatants sometimes not be killed? The answer is that rights do not always tell the whole moral story. Pity, empathy, or mercy are sometimes strong motivators not to kill. Sometimes (human) agents might be permitted to kill, but it might still be morally desirable for them not to do so. This argument does not depend on an account of human dignity. Rather, it articulates the common-sense view that killing is rarely morally desirable even if it is morally permissible. This is especially true during armed conflict where the designation of combatant status is sufficient to establish liability to attack. Often, as noted above, combatants are killed simply because they are in the wrong place at the wrong time, without having done anything.

The second challenge to the Argument from Human Agency is that it delivers too little too late. As the example of high-altitude bombing discussed earlier showed, modern combat technology has already distanced individuals from theatres in ways that make revising a decision to kill difficult. The difference, though, between more established weapons and out-of-the-loop systems is that the latter systems remove human agency entirely once the system has been deployed. Even in the case of high-altitude bombing, the operator has to decide whether to 'push the button'. Or, in the case of an on-the-loop system, the operator can override the systems' attack on a target. Granted; in reality, an operator's ability to override an on-the-loop system might be vanishingly small. If that is the case, there might be, as the Argument from Human Agency would concede, fewer reasons to think that AWS were morally unique. Rather, from the perspective of the Argument from Human Agency, many established forms of combat technology are more morally problematic than commonly assumed.

The third challenge is a more technical one for moral philosophy. If, according to the Argument from Human Agency, not killing is not strictly morally required because killing an enemy combatant via an AWS does not violate any moral obligations owed to that combatant, there could be strong reasons in favour of overriding the Argument from Human Agency. This would especially be the case when the deployment of AWS, as their defenders claim, led to significant reductions in casualties. Here, the Argument from Human Agency is weaker than dignity-based objections to AWS. In non-consequentialist or deontological moral theory, any trade-offs between beneficial aggregate consequences and dignity would be impermissible. The Argument from Human Agency, though, does not frame the issue in terms of human dignity. There might, thus, be some permissible trade-offs between human agency (deployment of human soldiers), on the one hand, and the aggregate number of lives saved via the deployment

[26] Leveringhaus, *Ethics and Autonomous Weapons* (n 11) 89–117.

of AWS, on the other. Still, the Argument from Human Agency illustrates that there is some loss when human agency is replaced with artificial agency. And that loss needs to clear a high justificatory bar. Here, the burden of proof falls on defenders of AWS.

To conclude, while the Argument from Human Agency captures intuitions about autonomous killing, it is not sufficient to show that it is categorically impermissible to replace human with artificial agency. It merely tries to raise the justificatory bar for AWS. The humanitarian gains from AWS must be high for the replacement of human agency with artificial agency to be morally legitimate. More generally, neither of the three positions examined above – the responsibility gap, human dignity, and human agency – serve as knockdown arguments against AWS. This is partly because, upon closer inspection, AWS are not more (or less) morally repugnant than established, and more accepted, weapons and associated forms of killing in war. In this light, it makes sense to shift the focus from the highly idealised scenario of AWS being deliberately programmed to attack human targets to different, and arguably more realistic, scenarios. Perhaps these alternative scenarios provide a clue as to why AWS might be morally problematic. The fourth and final part of the chapter looks at these scenarios in detail.

IV. THREE EMERGING ETHICAL PROBLEMS WITH AWS

As was emphasised earlier, for technological reasons, it is hard to see that the intentional programming of AWS in order to target combatants could be morally (or legally) permissible. As a result, the intended killing of combatants via AWS is not the main ethical challenge in the real world of AWS. Rather, AWS will be programmed to attack targets that are more easily and reliably identifiable by a machine. It is not far-fetched, for instance, to imagine an autonomous submarine that hunts other submarines, or an autonomous stealth plane programmed to fly into enemy territory and destroy radar stations, or a robot that can detect and eliminate enemy tanks. While these types of AWS are not deliberately programmed to attack human individuals, they still raise important ethical issues. In what follows, I focus on three of these.

First, the availability of AWS, some critics argue, has the potential to lead to more wars. Surely, in light of the destruction and loss of life that armed conflicts entail, this is a reason against AWS. If anything, we surely want fewer wars, not more. Yet, in the absence of counterfactuals, it is difficult to ascertain whether a particular form of weapons technology necessarily leads to more wars. If, for instance, the Soviet Union and US had not had access to nuclear weapons, would they have gone to war after 1945? It is impossible to tell. Moreover, it is noteworthy that a mere increase in armed conflict does not tell us anything about the justness of the resulting conflicts. Of course, if the availability of AWS increased the willingness of states to violate *jus ad bellum* by pursuing unjust wars, then these weapons are not normatively desirable. If, by contrast, the effect of AWS on the frequency of just or unjust wars was neutral, or if they increased the likelihood of just wars, they would, *ceteris paribus*, not necessarily be morally undesirable.

Yet, while it is not self-evident that AWS lead to an increase in unjust wars, their availability potentially lends itself to more covert and small-scale uses of force. Since the US's targeted killing campaign against suspected terrorists in the late 2000s, just war theorists have increasingly been concerned with uses of force that fall below the threshold for war and thus outside the regulatory frameworks provided *jus ad bellum* and *jus in bello*. Using the US-led War on Terror as a template, force is often used covertly and on an *ad hoc* basis, be it through the deployment of special forces or the targeting of alleged terrorists via remote-controlled aerial vehicles ('drones'), with few opportunities for public scrutiny and accountability. AWS might be ideal

for missions that are intended to fall, literally, under the radar. Once deployed, an AWS in stealth mode, without the need for further communication with a human operator, could enter enemy territory undetected and destroy a particular target, such as a military installation, a research facility, or even dual-use infrastructure. Although AWS should not be treated differently from other means used in covert operations, they may reinforce trends towards them.

Second, there is an unnerving analogy between AWS, landmines, and unexploded munitions, which often cause horrific damage in post-war environments. As we just saw, AWS can operate stealthily and without human oversight. With no direct human control over AWS, it is unclear how AWS can be deactivated after hostilities have been concluded. Rather unsettlingly, AWS, compared to landmines and unexploded munitions, could retain a much higher level of combat readiness. The moral issue is trivial and serious at the same time: does the very presence of autonomy in a weapon and the fact that it is an out-of-the-loop system make it difficult to switch it off? In other words, the central question is how, once human control over a weapon is ceded, it can be reasserted. How, for example, can a human operator re-establish control over an autonomous submarine operating in an undisclosed area of the high seas? There might eventually be technological answers to this question. Until then, the worry is that AWS remain a deadly legacy of armed conflict.

Third, while just war theorists have invested considerable energy into disambiguating the distinction between intended harm and unintended but foreseen harm, unintended and unforeseen harms, emanating from accidents and other misapplications of force, have received less attention. These harms are more widespread than assumed, leading to significant losses of life among non-combatants. Naturally, the fact that harm is unintended and unforeseen does not render it morally unproblematic. To the contrary, it raises questions about negligence and recklessness in armed conflict. One hypothesis in this respect, for instance, is that precision-weaponry has engendered reckless behaviour among belligerents.[27] Because these weapons are seen as precise, belligerents deploy them in high-risk theatres where accidents and misapplications of force are bound to happen. Here, abstention or the use of non-military alternatives seem more appropriate. For example, the use of military-grade weaponry, even if it is precise, over densely populated urban areas is arguably so risky that it is morally reckless. Belligerents know the risks but go ahead anyway because they trust the technology.

The conceptual relationship between precision-weaponry and AWS is not straightforward, but the question of recklessness is especially pertinent in the case of AWS.[28] After all, AWS not only create a significant kinetic effect, but they are unpredictable in doing so. As the saying goes, accidents are waiting to happen. True, in some cases, it might not be reckless to deploy AWS – for example, in extremely remote environments. But in many instances, and especially in the kinds of environments in which states have been conducting military operations over the last twenty-five years, it is morally reckless to deploy an inherently unpredictable weapon. Even if such a weapon is not deliberately programmed to directly attack human individuals, the threat it poses to human life is all too real. Can it really be guaranteed that an autonomous tank will not run over a civilian when speeding towards its target? What assurances can be given that an autonomous submarine does not mistake a boat carrying refugees for an enemy vessel? How can we be certain that a learning mechanism in a robotic weapon's governing software does not 'learn' that because a child once threw a rock at the robot during a military occupation, children in general constitute threats and should therefore be targeted? These worries are compounded

[27] B Cronin, *Bugsplat: The Politics of Collateral Damage in Western Armed Conflict* (2018).
[28] A Leveringhaus, 'Autonomous Weapons and the Future of Armed Conflict', in J Gailliot, D McIntosh, and JD Ohlin (eds), *Lethal Autonomous Weapons: Re-examining the Law and Ethics of Robotic Warfare* (2021) 175.

by the previous point about re-establishing control over an AWS. After control is ceded, it is not clear how it can be re-established, especially when it becomes apparent that the system does not operate in the way it should.

Advocates of AWS could mount two replies here. First, eventually there will be technological solutions that reduce the risk of accidents. Ultimately, this necessitates a technological assessment that ethicists cannot provide. The burden of proof, though, lies with technologists. Second, humans, defenders of AWS could point out, are also unpredictable, as the occurrence of war crimes or reckless behaviour during armed conflict attests. But the reply has three flaws. The first is that AWS will not be capable of offering a like-for-like replacement for human soldiers in armed conflict, especially when it comes to operations where the targets are enemy combatants (who would need to be differentiated from non-combatants). In this sense, the scope for human error, as well as wrongdoing, in armed conflict remains unchanged. The second flaw is that, although human individuals are unquestionably error-prone and unpredictable, AWS are unlikely, at the present stage of technological development, to perform any better than humans. The final flaw in the response is that, in the end, a fully armed weapons system has the capacity to do far more damage than any single soldier. For this reason alone, the deployment of AWS is, with few exceptions, morally reckless.

Taking stock, even if one turns from the highly abstract debate on AWS in contemporary philosophy to a more realistic appreciation of these weapons, moral problems and challenges do not magically disappear. Far from it, AWS potentially reinforce normatively undesirable dynamics in contemporary armed conflict, not least the push towards increasingly covert operations without public scrutiny, as well as the tendency for high-tech armies to (sometimes) take unreasonable, if not reckless, risks during combat operations. The key question of how control can be re-established over an out-of-the-loop system has not been satisfactorily answered, either. While these observations may not render AWS morally distinctive, they illustrate their *prima facie* undesirability.

V. CONCLUSION

Perhaps more than any other form of emerging weapons technology, AWS have been met with moral condemnation. As the analysis in this chapter shows, it is hard to pin down why they should be 'morally repugnant'. Some of the central ethical arguments against AWS do not withstand critical scrutiny. In particular, they fail to show that AWS are morally different from more established weapons and methods of warfighting. Still, the chapter concludes that AWS are morally problematic, though not necessarily morally repugnant. The main point here is that, for the foreseeable future, AWS are not safe enough to operate in what is often a complex and chaotic combat environment. This is not to say that their technological limitations might not eventually be overcome. But for now, the deployment of a weapon whose behaviour is to some extent unpredictable, without sufficient and on-going human oversight and the ability to rapidly establish operator control over it, seems morally reckless. True, other types of weapons can be used recklessly in armed conflict, too. The difference is that the technology underpinning AWS remains inherently unpredictable, and not just the use of these weapons. Furthermore, while AWS do not appear to raise fundamentally new issues in armed conflict, they seem to reinforce problematic dynamics in the use of force towards ever more covert missions. AWS might make it considerably easier for governments to avoid public scrutiny over their uses of force. Hence, for democratic reasons, and not just ethical ones, the arrival of AWS and the prospect of autonomous war fighting should be deeply troubling.

28

On 'Responsible AI' in War

Exploring Preconditions for Respecting International Law in Armed Conflict

Dustin A. Lewis

I. INTRODUCTION

In this chapter, I seek to help strengthen cross-disciplinary linkages in discourse concerning 'responsible Artificial Intelligence (AI)'. To do so, I explore certain aspects of international law pertaining to uses of AI-related tools and techniques in situations of armed conflict.

At least five factors compel increasingly urgent consideration of these issues by governments, scientists, engineers, ethicists, and lawyers, among many others. One aspect concerns the nature and the growing complexity of the socio-technical systems through which these technologies are configured. A second factor relates to the potential for more frequent – and possibly extensive – use of these technologies in armed conflicts. Those applications may span such areas as warfighting, detention, humanitarian services, maritime systems, and logistics. A third issue concerns potential challenges and opportunities concerning the application of international law to employments of AI-related tools and techniques in armed conflicts. A fourth dimension relates to debates around whether or not the existing international legal framework applicable to armed conflicts sufficiently addresses ethical concerns and normative commitments implicated by AI – and, if it does not, how the framework ought to be adjusted. A fifth element concerns a potential 'double black box' in which humans encase technical opacity in military secrecy.

One way to seek to help identify and address potential issues and concerns in this area is to go 'back to the basics' by elaborating some key elements underpinning legal compliance, responsibility, and agency in armed conflict. In this chapter, I aim to help illuminate some of the preconditions arguably necessary for respecting international law with regard to employments of AI-related tools and techniques in armed conflicts. By respecting international law, I principally mean two things: (1) applying and observing international law with regard to relevant conduct and (2) facilitating incurrence of responsibility for violations arising in connection with relevant conduct. (The latter might be seen either as an integral element or a corollary of the former.) Underlying my exploration is the argument that there may be descriptive and normative value in framing part of the discussion related to 'responsible AI' in terms of discerning and instantiating the preconditions necessary for respecting international law.

I proceed as follows. In Section II, I frame some contextual aspects of my inquiry. In Section III, I sketch a brief primer on international law applicable to armed conflict. In Section IV, I set out some of the preconditions arguably necessary to respect international law. In Section V, I briefly conclude.

Two caveats ought to be borne in mind. The first caveat is that the bulk of the research underlying this chapter drew primarily on English-language materials. The absence of a broader examination of legal materials, scholarship, and other resources in other languages narrows the study's scope. The second caveat is that this chapter seeks to set forth, in broad-brush strokes, some of the preconditions arguably underpinning respect for international law.[1] Therefore, the analysis and the identification of potential issues and concerns are far from comprehensive. Analysis in respect of particular actors, armed conflicts, or AI-related tools and techniques may uncover (perhaps numerous) additional preconditions.

II. FRAMING

In this section, I frame some contextual aspects of my inquiry. In particular, I briefly outline some elements concerning definitions of AI. I also enumerate some existing and anticipated uses for AI in armed conflict. Next, I sketch the status of international discussions on certain military applications of possibly related technologies. And, finally, I highlight issues around technical opacity combined with military secrecy.

1. *Definitional Parameters*

Terminological inflation may give rise to characterizations of various technologies as 'AI' even where those technologies do not fall into recognized definitions of AI. Potentially complicating matters further is that there is no agreed definition of AI expressly laid down in an international legal instrument applicable to armed conflict.

For this chapter, I will assume a relatively expansive definition of AI, one drawn from my understanding – as a non-scientific-expert – of AI science broadly conceived.[2] It may be argued that AI science pertains in part to the development of computationally-based understandings of intelligent behaviour, typically through two interrelated steps. One step relates to the determination of cognitive structures and processes and the corresponding design of ways to represent and reason effectively. The other step concerns developing (a combination of) theories, models, data, equations, algorithms, or systems that 'embody' that understanding. Under this approach, AI systems are sometimes conceived as incorporating techniques or using tools that enable systems to 'reason' more or less 'intelligently' and to 'act' more or less 'autonomously.' The systems might do so by, for example, interpreting natural languages and visual scenes; 'learning' (in the sense of training); drawing inferences; or making 'decisions' and taking action on those 'decisions'. The techniques and tools might be rooted in one or more of the following

[1] My analysis in this chapter – and especially Section IV – draws heavily on, and reproduces certain text from, a DA Lewis, 'Preconditions for Applying International Law to Autonomous Cyber Capabilities', in R Liivoja and A Väljataga (eds), *Autonomous Cyber Capabilities under International Law* (NATO Cooperative Cyber Defence Centre of Excellence, 2021). Both the current chapter and that piece draw on the work of a research project at the Harvard Law School Program on International Law and Armed Conflict titled 'International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns' (Harvard Law School Program on International Law and Armed Conflict, 'Project on International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns' (November 2019) https://pilac.law.harvard.edu/international-legal-and-policy-dimensions-of-war-algorithms). That project seeks to strengthen international debate and inform policy-making on the ways that AI and complex computer algorithms are transforming, and have the potential to reshape, war.

[2] This paragraph draws extensively on DA Lewis, 'Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider' (*ICRC Humanitarian Law and Policy Blog*, 21 March 2019) https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/ (hereafter Lewis, 'Legal Reviews'); see also W Burgard, Chapter 1, in this volume.

methods: those rooted in logical reasoning broadly conceived, which are sometimes also referred to as 'symbolic AI' (as a form of model-based methods); those rooted in probability (also as a form of model-based methods); or those rooted in statistical reasoning and data (as a form of data-dependent or data-driven methods).

2. *Diversity of Applications*

Certain armed forces have long used AI-related tools and techniques. For example, in relation to the Gulf War of 1990–91, the United States employed a program called the Dynamic Analysis and Replanning Tool (DART), which increased efficiencies in scheduling and making logistical arrangements for the transportation of supplies and personnel.[3]

Today, existing and contemplated applications of AI-related tools and techniques related to warfighting range widely.[4] With the caveat concerning terminological inflation noted above in mind, certain States are making efforts to (further) automate targeting-related communications support,[5] air-to-air combat,[6] anti-unmanned-aerial-vehicle countermeasures,[7] so-called loitering-attack munitions,[8] target recognition,[9] and analysis of intelligence, reconnaissance, and surveillance sources.[10] Armed forces are developing machine-learning techniques to generate

[3] See M Bienkowski, 'Demonstrating the Operational Feasibility of New Technologies: the ARPI IFDs' (1995) 10(1) *IEEE Expert* 27, 28–29.

[4] See, e.g., MAC Ekelhof and G Paoli, 'The Human Element in Decisions about the Use of Force' (UN Institute for Disarmament Research, 2020) https://unidir.org/publication/human-element-decisions-about-use-force; E Kania, '"AI Weapons" in China's Military Innovation' (Brookings Institution, April 2020) www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf; MAC Ekelhof and GP Paoli, 'Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems' (2020) UN Institute for Disarmament Research https://unidir.org/sites/default/files/2020-04/UNIDIR_Swarms_SinglePages_web.pdf; MAC Ekelhof, 'The Distributed Conduct of War: Reframing Debates on Autonomous Weapons, Human Control and Legal Compliance in Targeting' (PhD Dissertation, Vrije Universiteit 2019); KM Sayler, 'Artificial Intelligence and National Security' (21 November 2019) Congressional Research Service Report No R45178 https://fas.org/sgp/crs/natsec/R45178.pdf; International Committee of the Red Cross, 'Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control' (ICRC Report, August 2019) www.icrc.org/en/download/file/102852/autonomy_artificial_intelligence_and_robotics.pdf; United Nations Institute for Disarmament Research, 'The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence – A Primer for CCW delegates' (2018) UNIDIR Paper No 8 https://unidir.org/publication/weaponization-increasingly-autonomous-technologies-artificial-intelligence; MAC Ekelhof, 'Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control the Lens of Military Targeting' (2018) 73 *Nav War Coll Rev* 61; P Sharre, *Army of One* (2018) 27–56; V Boulanin and M Verbruggen, 'Mapping the Development of Autonomy in Weapons Systems' (Stockholm International Peace Research Institute, 2017) www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf.

[5] 'DOD Official Briefs Reporters on Artificial Intelligence Developments' (Transcript of Nand Mulchandani, 8 July 2020) www.defense.gov/Newsroom/Transcripts/Transcript/Article/2270329/dod-official-briefs-reporters-on-artificial-intelligence-developments/.

[6] K Reichmann, 'Can Artificial Intelligence Improve Aerial Dogfighting?' (*C4ISRNET*, 7 June 2019) www.c4isrnet.com/artificial-intelligence/2019/06/07/can-artificial-intelligence-improve-aerial-dogfighting/.

[7] See Industry News Release, 'Air Force to Deploy Citadel Defense Titan CUAS Solutions to Defeat Drone Swarms' *Defense Media Network* (17 September 2019) www.defensemedianetwork.com/stories/air-force-to-deploy-citadel-defense-titan-cuas-solutions-to-defeat-drone-swarms/.

[8] See, e.g., D Gettinger and AH Michel, 'Loitering Munitions' (Center for the Study of the Drone, 2 February 2017) https://dronecenter.bard.edu/files/2017/02/CSD-Loitering-Munitions.pdf.

[9] On legal aspects of automatic target recognition systems involving 'deep learning' methods, see JG Hughes, 'The Law of Armed Conflict Issues Created by Programming Automatic Target Recognition Systems Using Deep Learning Methods' (2018) 21 *YBIHL* 99.

[10] See, e.g., N Strout, 'Inside the Army's Futuristic Test of Its Battlefield Artificial Intelligence in the Desert' (*C4ISRNET*, 25 September 2020) www.c4isrnet.com/artificial-intelligence/2020/09/25/the-army-just-conducted-a-massive-test-of-its-battlefield-artificial-intelligence-in-the-desert/.

targeting data.[11] Prototypes of automated target-recognition heads-up displays are also under development.[12] Rationales underlying these efforts are often rooted in military doctrines and security strategies that place a premium on enhancing speed and agility in decision-making and tasks and preserving operational capabilities in restricted environments.[13]

In the naval context, recent technological developments – including those related to AI – afford uninhabited military maritime systems, whether on or below the surface, capabilities to navigate and explore with less direct ongoing human supervision and interaction than before. Reportedly, for example, China is developing a surface system called the JARI that, while remotely controlled, purports to use AI to autonomously navigate and undertake combat missions once it receives commands.[14]

The likelihood seems to be increasing that AI-related tools and techniques may be used to help make factual determinations as well as related evaluative decisions and normative judgements around detention in armed conflict.[15] Possible antecedent technologies include algorithmic filtering of data and statistically-based risk assessments initially created for domestic policing and criminal-law settings. Potential applications in armed conflict might include prioritizing military patrols, assessing levels and kinds of threats purportedly posed by individuals or groups, and determining who should be held and when someone should be released. For example, authorities in Israel have reportedly used algorithms as part of attempts to obviate anticipated attacks by Palestinians through a process that involves the filtering of social-media data, resulting in over 200 arrests.[16] (It is not clear whether or not the technologies used in that context may be characterized as AI.)

It does not seem to strain credulity to anticipate that the provision of humanitarian services in war – both protection and relief activities[17] – may rely in some contexts on AI-related tools and techniques.[18] Applications that might be characterized as relying on possible technical antecedents to AI-related tools and techniques include predictive-mapping technologies used to inform populations of outbreaks of violence, track movements of armed actors, predict population movements, and prioritize response resources.[19]

[11] See N Strout, 'How the Army Plans to Use Space and Artificial Intelligence to Hit Deep Targets Quickly' *Defense News* (5 August 2020) www.defensenews.com/digital-show-dailies/smd/2020/08/05/how-the-army-plans-to-use-space-and-artificial-intelligence-to-hit-deep-targets-quickly/.

[12] See J Keller, 'The Army's Futuristic Heads-Up Display Is Coming Sooner than You Think' (*Task & Purpose*, 20 November 2019) https://taskandpurpose.com/military-tech/army-integrated-visual-augmentation-system-fielding-date.

[13] See CP Trumbull IV, 'Autonomous Weapons: How Existing Law Can Regulate Future Weapons' (2020) 34 *EmoryILR* 533, 544–550.

[14] See L Xuanzun, 'China Launches World-Leading Unmanned Warship' Global Times (22 August 2019) www.globaltimes.cn/content/1162320.shtml.

[15] See DA Lewis, 'AI and Machine Learning Symposium: Why Detention, Humanitarian Services, Maritime Systems, and Legal Advice Merit Greater Attention' (*Opinio Juris*, 28 April 2020) http://opiniojuris.org/2020/04/28/ai-and-machine-learning-symposium-ai-in-armed-conflict-why-detention-humanitarian-services-maritime-systems-and-legal-advice-merit-greater-attention/ (hereafter Lewis, 'AI and Machine Learning'); T Bridgeman, 'The Viability of Data-Reliant Predictive Systems in Armed Conflict Detention' (*ICRC Humanitarian Law and Policy Blog*, 8 April 2019) https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention/ ; A Deeks, 'Detaining by Algorithm' (*ICRC Humanitarian Law and Policy Blog*, 25 March 2019) https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/; A Deeks, 'Predicting Enemies' (2018) 104 Virginia LR 1529.

[16] CBS News, 'Israel Claims 200 Attacks Predicted, Prevented with Data Tech' *CBS News* (12 June 2018) www.cbsnews.com/news/israel-data-algorithms-predict-terrorism-palestinians-privacy-civil-liberties/.

[17] See ICRC, *Commentary on the First Geneva Convention: Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field* (2nd ed. 2016) paras 807–821 https://ihl-databases.icrc.org/ihl/full/GCI-commentary (hereafter ICRC, *Commentary*).

[18] See Lewis, 'AI and Machine Learning' (n 15).

[19] See UNHCR, 'The Jetson Story' (UN High Commissioner for Refugees Innovation Service) http://jetson.unhcr.org/story.html; N Manning, 'Keeping the Peace: The UN Department of Field Service's and Peacekeeping Operations Use of Ushahidi' (*Ushahidi Blog*, 8 August 2018) www.ushahidi.com/blog/2018/08/08/keeping-the-peace-the-un-depart

3. *International Debates on 'Emerging Technologies in the Area of Lethal Autonomous Weapons Systems'*

Perhaps especially since 2013, increased attention has been given at the international level to issues around autonomous weapons. Such weapons may or may not involve AI-related tools or techniques. A significant aspect of the debate appears to have reached a kind of normative deadlock.[20] That impasse has arisen in the recent main primary venue for intergovernmental discourse: the Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapons systems (GGE), which was established under the Convention on Certain Conventional Weapons (CCW)[21] in 2016.

GGE debates on the law most frequently fall under three general categories: international humanitarian law/law of armed conflict (IHL/LOAC) rules on the conduct of hostilities, especially on distinction, proportionality, and precautions in attacks; reviews of weapons, means, and methods of warfare;[22] and individual and State responsibility.[23] (The primary field of international law developed by States to apply to conduct undertaken in relation to armed conflict is now often called IHL/LOAC; this field is sometimes known as the *jus in bello* or the laws of war.)

Perhaps the most pivotal axis of the current debate concerns the desirability (or not) of developing and instantiating a concept of 'meaningful human control' or a similar formulation over the use of force, including autonomy in configuring, nominating, prioritizing, and applying force to targets.[24] A close reading of States' views expressed in the GGE suggests that

ment-of-field-services-and-peacekeeping-operations-use-of-ushahidi. See also A Duursma and J Karlsrud, 'Predictive Peacekeeping: Strengthening Predictive Analysis in UN Peace Operations' (2019) 8 *Stability IJ Sec & Dev* 1.

[20] This section draws heavily on DA Lewis, 'An Enduring Impasse on Autonomous Weapons' (*Just Security*, 28 September 2020) www.justsecurity.org/72610/an-enduring-impasse-on-autonomous-weapons/ (hereafter Lewis, 'An Enduring Impasse').

[21] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (with Protocols I, II, and III) (signed 10 October 1980, entry into force 2 December 1983) 1342 UNTS 137.

[22] See GGE, 'Questionnaire on the Legal Review Mechanisms of New Weapons, Means and Methods of Warfare' (29 March 2019) Working Paper by Argentina to the Group of Governmental Experts on Lethal Autonomous Weapons Systems CCW/GGE.1/2019/WP.6; GGE, *'The Australian Article 36 Review Process' (30 August 2018) Working Paper by Australia to the* Group of Governmental Experts on Lethal Autonomous Weapons Systems CCW/GGE.2/2018/WP6; GGE, *'Strengthening of the Review Mechanisms of a New Weapon, Means or Methods of Warfare' (4 April 2018) Working Paper by Argentina to the Group of Governmental Experts on Lethal Autonomous Weapons Systems* CCW/GGE.1/2018/WP2; GGE, *'Weapons Review Mechanisms' (7 November 2017) Working Paper by the Netherlands and Switzerland to the Group of Governmental Experts on Lethal Autonomous Weapons Systems* CCW/GGE.1/2017/WP5; German Defense Ministry, 'Statement on the Implementation of Weapons Reviews under Article 36 Additional Protocol I by Germany' (The Convention on Certain Conventional Weapons (CCW) Third Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 11–15 April 2016) https://perma.cc/4EFG-LCEM; M Meier, 'US Delegation Statement on "Weapon Reviews"' (The Convention on Certain Conventional Weapons (CCW) Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13 April 2016) www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2016/meeting-experts-laws/statements/13April_US.pdf.

[23] M Brenneke, 'Lethal Autonomous Weapon Systems and Their Compatibility with International Humanitarian Law: A Primer on the Debate' (2018) 21 *YBIHL* 59.

[24] See M Wareham 'Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control' *(Human Rights Watch, August 2020)* www.hrw.org/sites/default/files/media_2020/08/arms0820_web.pdf; AM Eklund, 'Meaningful Human Control of Autonomous Weapon Systems: Definitions and Key Elements in the Light of International Humanitarian Law and International Human Rights Law' (Swedish Defense Research Agency FOI, February 2020) www.fcas-forum.eu/publications/Meaningful-Human-Control-of-Autonomous-Weapon-Systems-Eklund.pdf; V Boulanin and others, 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control' (Stockholm International Peace Research Institute and International Committee of

governments hold seemingly irreconcilable positions beyond some generically formulated principles, at least so far, on whether existing law is fit for purpose or new law is warranted.[25] That said, there might be a large enough contingent to pursue legal reform, perhaps outside of the CCW.

4. *Technical Opacity Coupled with Military Secrecy*

Both inside and outside of the GGE, armed forces continue to be deeply reluctant to disclose how they configure sensors, algorithms, data, and machines, including as part of their attempts to satisfy legal rules applicable in relation to war. In a nutshell, a kind of 'double black box' may emerge where human agents encase technical opacity in military secrecy.[26]

The specific conduct of war as well as military-technological capabilities are rarely revealed publicly by States and non-state parties to armed conflicts. Partly because of that, it is difficult for people outside of armed forces to reliably discern whether new technological affordances create or exacerbate challenges (as critics allege) or generate or amplify opportunities (as proponents assert) for greater respect for the law and more purportedly 'humanitarian' outcomes.[27] It is difficult to discern, for example, how and to what extent the human agents composing a party to an armed conflict in practice construct and correlate proxies for legally relevant characteristics – for example, those concerning direct participation in hostilities as a basis for targeting[28] or imperative reasons of security as a ground for detention[29] – involved in the collection of data and the operation of algorithms. Nor do parties routinely divulge what specific dependencies exist within and between the computational components that their human agents adopt regarding a particular form of warfare. Instead, by and large, parties – at most – merely reaffirm in generic terms that their human agents strictly respect the rules.

the Red Cross, June 2020) www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy_0.pdf (hereafter Boulanin and others, 'Limits on Autonomy'); ICRC, 'Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach' (International Committee of the Red Cross, 6 June 2019) www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach; T Singer, *Dehumanisierung der Kriegführung: Herausforderungen für das Völkerrecht und die Frage nach der Notwendigkeit menschlicher Kontrolle* (2019); Advisory Council on International Affairs and Advisory Committee on Issues of Public International Law, *Autonomous Weapon Systems; the Need for Meaningful Control* (No. 97 AIV/ No. 26 CAVV, October 2015) (views adopted by Government) www.advisorycouncilinternationalaffairs.nl/documents/publications/2015/10/02/autonomous-weapon-systems; Working Paper by Austria, 'The Concept of "Meaningful Human Control"' (The Convention on Certain Conventional Weapons (CCW) Second Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13–18 April 2015) https://perma.cc/D35A-RP7G.

[25] See, e.g., Lewis, 'An Enduring Impasse' (n 20).

[26] See generally AH Michel, 'The Black Box, Unlocked: Predictability and Understandability in Military AI' (UN Institute for Disarmament Research, 2020) https://unidir.org/publication/black-box-unlocked (hereafter Michel, 'The Black Box, Unlocked').

[27] See, e.g., Lewis, 'An Enduring Impasse' (n 20).

[28] See Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) (signed 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 (Additional Protocol I) Art 51(3) (hereafter AP I); Protocol Additional to the Geneva Conventions of 12 August 1949 and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II) (signed 8 June 1977, entered into force 7 December 1978) 1125 UNTS 609 (Additional Protocol II) Article 13(3) (hereafter AP II).

[29] See Geneva Convention relative to the Protection of Civilian Persons in Time of War (signed 12 August 1949, entry into force 21 October 1950) 75 UNTS 287 (GC IV) Article 78, first para.

III. OVERVIEW OF INTERNATIONAL LAW APPLICABLE TO ARMED CONFLICT

International law is the only binding framework agreed to by States to regulate acts and omissions related to armed conflict. In this respect, international law is distinguishable from national legal frameworks, corporate codes of conduct, and ethics policies.

The sources, or origins, of international law applicable in relation to armed conflict include treaties, customary international law, and general principles of law. Several fields of international law may lay down binding rules applicable to a particular armed conflict. As mentioned earlier, the primary field developed by States to apply to conduct undertaken in relation to armed conflict is IHL/LOAC. Other potentially relevant fields may include the area of international law regulating the threat or use of force in international relations (also known as the *jus ad bellum* or the *jus contra bellum*), international human rights law, international criminal law, international refugee law, the law of State responsibility, and the law of responsibility of international organizations. In international law, an international organization (IO) is often defined as an organization established by a treaty or other instrument governed by international law and possessing its own international legal personality.[30] Examples of IOs include the United Nations Organization (UN) and the North Atlantic Treaty Organization (NATO), among many others.

Under contemporary IHL/LOAC, there are two generally recognized classifications, or categories, of armed conflicts.[31] One is an international armed conflict, and the other is a non-international armed conflict. The nature of the parties most often distinguishes these categories. International armed conflicts are typically considered to involve two or more States as adversaries. Non-international armed conflicts generally involve one or more States fighting together against one or more non-state parties or two or more non-state parties fighting against each other.

What amounts to a breach of IHL/LOAC depends on the content of the underlying obligation applicable to a particular human or legal entity. Depending on the specific armed conflict, potentially relevant legal entities may include one or more States, IOs, or non-state parties. IHL/LOAC structures and lays down legal provisions concerning such thematic areas as the conduct of hostilities, detention, and humanitarian services, among many others.

For example, under certain IHL/LOAC instruments, some weapons are expressly prohibited, such as poisoned weapons,[32] chemical weapons,[33] and weapons that injure by fragments that escape detection by X-rays in the human body.[34] The use of weapons that are not expressly prohibited may be tolerated under IHL/LOAC at least insofar as the use of the weapon comports with applicable provisions. For instance, depending on the specific circumstances of use and the relevant actors, those provisions may include:

[30] See Draft Articles on Responsibility of International Organizations with Commentary (Report of the Commission to the General Assembly on the Work of Its Sixty-Third Session, 2011) Ybk Intl L Comm, Volume II (Part 2) A/CN.4/SER.A/2011/Add 1 (Part 2), Article 2(a) (hereafter (D)ARIO).

[31] See ICRC, *Commentary* (n 17) paras 201–342, 384–502.

[32] Regulations Respecting the Laws and Customs of War on Land, Annex to Convention (IV) Respecting the Laws and Customs of War on Land (signed 18 October 1907, entered into force 26 January 1910) 36 Stat 2295, Article 23(a).

[33] Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction (signed 3 September 1992, entered into force 29 April 1997) 1975 UNTS 45, Article I(1).

[34] Protocol on Non-detectable Fragments (Protocol I) to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (signed 10 October 1980, entered into force 2 December 1983) 1342 UNTS 147.

- the obligation for parties to distinguish between the civilian population and combatants and between civilian objects and military objectives and to direct their operations only against military objectives;[35]
- the prohibition on attacks which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated;[36]
- the obligation to take constant care to spare the civilian population, civilians, and civilian objects in military operations;[37] and
- obligations to take certain precautions concerning attacks.[38]

International law sets out particular standard assumptions of responsibility for the conduct of States and IOs. It is on the basis of those assumptions that specific IHL/LOAC provisions exist and are applied.[39] In other words, international law pertaining to armed conflict exists and is applied in respect of States and IOs based on the interrelationships between the 'primary' substantive IHL/LOAC provisions and the 'secondary' responsibility institutions. Regarding both State responsibility and IO responsibility, standard assumptions of responsibility are rooted in underlying concepts of attribution, breach, circumstances precluding wrongfulness, and consequences.[40] Those assumptions are general in character and are assumed and apply unless excluded, for example through an individual treaty or rule.[41]

A use in an armed conflict of an AI-related tool or technique may (also or separately) give rise to individual criminal responsibility under international law. Such personal criminal responsibility may arise where the conduct that forms the application of an AI-related tool or technique constitutes, or otherwise sufficiently contributes to, an international crime. For example, under the Rome Statute of the International Criminal Court (ICC), the court has jurisdiction over the crime of genocide, crimes against humanity, war crimes, and the crime of aggression.[42] A use of an AI-related tool or technique may form part or all of the conduct underlying one or more of the crimes prohibited under the ICC Statute.

Concerning imposition of individual criminal responsibility, it may be argued that standard assumptions of responsibility are based (at least under the ICC Statute) on certain underlying concepts.[43] Those concepts may arguably include jurisdiction;[44] ascription (that is, attribution of conduct to a natural person);[45] material elements (in the sense of the prohibited conduct forming the crime);[46] mental elements (including the requisite intent and knowledge);[47] modes

[35] AP I (n 28) Article 48.
[36] Ibid Article 51(5)(*b*).
[37] Ibid Article 57(1).
[38] Ibid Article 57(2).
[39] See JR Crawford, 'State Responsibility' in R Wolfrum (ed), *Max Planck Encyclopedia of Public International Law* (2006) (hereafter Crawford, 'State Responsibility').
[40] Ibid; Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentary (Report of the Commission to the General Assembly on the Work of its Fifty-Third Session, 2001) Ybk Intl L Comm, Volume II (Part Two) A/CN.4/SER.A/2001/Add 1 (Part 2) (hereafter (D)ARSIWA); (D)ARIO (n 30).
[41] Crawford, 'State Responsibility' (n 39).
[42] Rome Statute of the International Criminal Court (signed 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3 (ICC Statute), Articles 5, 10–19.
[43] See DA Lewis, 'International Legal Regulation of the Employment of Artificial-Intelligence-Related Technologies in Armed Conflict' (2020) 2 *Moscow JIL* 53, 61–63.
[44] See ICC Statute, Articles 5–19.
[45] See ICC Statute, Articles 25–26.
[46] See ICC Statute, Articles 6–8 *bis*.
[47] See ICC Statute, Article 30.

of responsibility (such as aiding and abetting or command responsibility);[48] grounds for excluding responsibility;[49] trial;[50] penalties (including imprisonment of the responsible person);[51] and appeal and revision.[52] It may be argued that it is on the basis of the assumptions related to those concepts that the provisions of the ICC Statute exist and are applied.

IV. PRECONDITIONS ARGUABLY NECESSARY TO RESPECT INTERNATIONAL LAW

In this section, I outline some preconditions underlying elements that are arguably necessary for international law to be respected in relation to a use in an armed conflict of an AI-related tool or technique. I assume that the employment of the technology is governed (at least in part) by international law. By respecting international law, I mean the bringing of a binding norm, principle, rule, or standard to bear in relation to a particular employment of an AI-related tool or technique in a manner that accords with the object and purpose of the relevant provision, that facilitates observance of the provision, and that facilitates incurrence of responsibility in case of breach of the provision.

At least three categories of actors may be involved in respecting international law in relation to a use in an armed conflict of an AI-related tool or technique. Each category is arguably made up, first and foremost, of human agents. In addition to those human agents, the entities to which those humans are attached or through which they otherwise (seek to) implement international law may also be relevant.

The first category is made up in part of the humans who are *involved* in relevant acts or omissions (or both) that form the employment of an AI-related tool or technique attributable to a State or an IO. This first category of actors also includes the entity or entities – such as the State or the IO or some combination of State(s) and IO(s) – to which the employment is attributable. The human agents may include, for example, software engineers, operators, commanders, and legal advisers engaging in conduct on behalf of the State or the IO.

The second category of actors is made up in part of humans *not involved* in the employment in an armed conflict of an AI-related tool or technique attributable to a State or an IO but who may nevertheless (seek to) ensure respect for international law in relation to that conduct. This second category of actors also includes entities – such as (other) States, (other) IOs, international courts, and the like – that may attempt, functionally through the humans who compose them, to ensure respect for international law in relation to the conduct.

The third category of actors is made up in part of humans who (seek to) apply international law – especially international law on international crimes – to relevant conduct of a natural person. These humans may include, for example, prosecutors, defense counsel, and judges. This third category of actors also includes entities (mostly, but not exclusively, international or domestic criminal tribunals) that may seek, functionally through the humans who compose them, to apply international law to natural persons.

In the rest of this section, I seek to elaborate some preconditions regarding each of these three respective categories of actors.

[48] See ICC Statute, Articles 25, 28.
[49] See ICC Statute, Articles 31–33.
[50] See ICC Statute, Articles 62–76.
[51] See ICC Statute, Article 77.
[52] ICC Statute, Articles 81–84.

1. *Preconditions Concerning Respect for International Law by Human Agents Acting on Behalf of a State or an International Organization*

In this sub-section, I focus on employments in armed conflicts of AI-related tools or techniques attributable to one or more States, IOs, or some combination thereof. In particular, I seek to outline some preconditions underlying elements that are arguably necessary for the State or the IO to respect international law in relation to such an employment.

Precondition #1: Humans Are Legal Agents of States and International Organizations
The first precondition is that humans are arguably the agents for the exercise and implementation of international law applicable to States and IOs. This precondition is premised on the notion that existing international law presupposes that the functional exercise and implementation of international law by a State or an IO in relation to the conduct of that State or that IO is reserved solely to humans.[53] According to this approach, this primary exercise and implementation of international law may not be partly or wholly reposed in non-human (artificial) agents.[54]

Precondition #2: Human Agents of the State or the International Organization Sufficiently Understand the Performance and Effects of the Employment
The second precondition is that human agents of the State or the IO that engages in conduct that forms an employment in an armed conflict of an AI-related tool or technique arguably need to sufficiently understand the technical performance and effects of the employed tool or technique in respect of the specific circumstances of the employment and in relation to the socio-technical system through which the tool or technique is employed.[55] For this precondition to be instantiated, the understanding arguably needs to encompass (among other things) comprehension of the dependencies underlying the socio-technical system, the specific circumstances and conditions of the employment, and the interactions between those dependencies, circumstances, and conditions.

[53] See Informal Working Paper by Switzerland (30 March 2016), 'Towards a "Compliance-Based" Approach to LAWS [Lethal Autonomous Weapons Systems]' (Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 11–15 April 2016) https://perma.cc/WRJ6-CCMS (expressing the position that '[t]he Geneva Conventions of 1949 and the Additional Protocols of 1977 were undoubtedly conceived with States and individual humans as agents for the exercise and implementation of the resulting rights and obligations in mind.') (hereafter Switzerland, 'Towards a "Compliance-Based" Approach'); see also Office of the General Counsel of the Department of Defense (US), *Department of Defense Law of War Manual* [June 2015, updated Dec. 2016], s 6.5.9.3, p 354 (expressing the position that law-of-war obligations apply to persons rather than to weapons, including that 'it is persons who must comply with the law of war') (hereafter US DoD OGC, *Law of War Manual*).

[54] For an argument that algorithmic forms of warfare – which may apparently include certain employments of AI-related tools or techniques – cannot be subject to law writ large, see G Noll, 'War by Algorithm: The End of Law?', in M Liljefors, G Noll, and D Steuer (eds), *War and Algorithm* (2019).

[55] See generally L Suchman, 'Configuration' in C Lury and N Wakeford (eds), *Inventive Methods* (2012). For an analysis of the 'technical layer', the 'socio-technical layer,' and the 'governance layer' pertaining to autonomous weapons systems, see I Verdiesen, F Santoni de Sio, and V Dignum, 'Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight'(2020) *Minds and Machines* https://doi.org/10.1007/s11023-020-09532-9. For an analysis of US 'drone operations' (albeit admittedly not pertaining to AI as such) informed in part by methods relevant to socio-technical configurations, see MC Elish, 'Remote Split: A History of US Drone Operations and the Distributed Labor of War' (2017) 42(6) *Science, Technology, & Human Values* 1100. On certain issues related to predicting and understanding military applications of artificial intelligence, see Michel, 'The Black Box, Unlocked' (n 26). With respect to machine-learning algorithms more broadly, see J Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (January–June 2016) *Big Data & Society* 1-12. For recent arguments concerning limits on autonomy in weapons systems in particular, see Boulanin and others, 'Limits on Autonomy' (n 24).

Precondition #3: Human Agents of the State or the International Organization Discern the Law Applicable to the Employment
The third precondition is that human agents of the State or the IO that engages in conduct that forms an employment in an armed conflict of an AI-related tool or technique arguably need to discern the law applicable to the State or the IO in relation to the employment. The applicable law may vary based on (among other things) the specific legal provisions applicable to the State or the IO through different sources, or origins, of international law. (As noted above, those sources may include treaty law, customary international law, and general principles of international law, among others.)

Precondition #4: Human Agents of the State or the International Organization Assess the Legality of the Anticipated Employment Before the Employment
The fourth precondition is that human agents of the State or the IO that engages in conduct that forms an employment in an armed conflict of an AI-related tool or technique assess – before the employment is initiated – whether the anticipated employment would conform with applicable law in relation to the anticipated specific circumstances and conditions of the employment.[56] In line with this precondition, only those employments that pass this legality assessment may be initiated and only then under the circumstances and subject to the conditions necessary to pass this legality assessment.

Precondition #5: Human Agents of the State or the International Organization Impose Legally Mandated Parameters Before and During the Employment
The fifth precondition is that human agents of the State or the IO that engages in conduct that forms an employment in an armed conflict of an AI-related tool or technique need to impose – before and during the employment – limitations or prohibitions or both as required by applicable law in respect of the employment. To instantiate this precondition, human agents of the State or the IO need to discern and configure the particular limitations or prohibitions by interpreting and applying international law in respect of the employment. Factors that the human agents might need to consider could include (among many others) interactions between the socio-technical system's dependencies and the specific circumstances and conditions of the employment.[57]

Suppose those dependencies, circumstances, or conditions (or some combination thereof) materially change after the employment is initiated. In that case, the human agents of the State or the IO arguably need to discern and configure the limitations or prohibitions (or both) in light of those changes.

To the extent, if any, required by the law applicable in relation to a specific employment or generally, human agents of the State or the IO may need to facilitate at least partial interaction by one or more humans with the system during the employment. Such interactions may take such forms (among others) as monitoring, suspension, or cancellation of some or all of the employment.[58]

[56] See N Goussac, 'Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting' (*ICRC Humanitarian Law and Policy Blog*, 18 April 2019) https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/; Lewis, 'Legal Reviews' (n 2).

[57] For broader critiques and concerns – including some informed by socio-technical perspectives – related to (over-) reliance on algorithmic systems, see, among others, R Benjamin, *Race after Technology* (2019); SU Noble, *Algorithms of Oppression* (2018); BD Mittelstadt and others, 'The Ethics of Algorithms: Mapping the Debate' (July–Dec. 2016) *Big Data & Society* 1-21; C O'Neil, *Weapons of Math Destruction* (2016).

[58] See, e.g., with respect to precautions in attacks in situations of armed conflict, AP I (n 28) Article 57(2)(b).

Precondition #6: Human Agents of the State or the International Organization Assess (Il)Legality after the Employment

The sixth precondition is that human agents of the State or the IO that engages in conduct that forms an employment in an armed conflict of an AI-related tool or technique arguably need to assess, after employment, whether or not the employment complied with applicable law. To instantiate this precondition, those human agents need to discern (among other things) which humans engaged in which elements of relevant conduct, the circumstances and conditions pertaining to that conduct, and whether the anticipated and actual performance and effects of the socio-technical system underlying the employment conformed with the legally mandated parameters.

Precondition #7: Human Agents of the State or the International Organization Assess Potential Responsibility for Violations Arising in Connection with the Employment

The seventh precondition concerns suspected violations that may arise in relation to an employment in an armed conflict of an AI-related tool or technique by or on behalf of a State or an IO. The precondition is that human agents of the State or the IO that undertook the conduct assess whether or not the conduct constitutes a violation – and, if they assess a violation occurred, human agents of the State or the IO (also) evaluate whether the international legal responsibility of the State or the IO is engaged. To make the assessment required by this precondition, human agents of the State or the IO need to discern, first, whether or not the conduct that forms the employment is attributable to the State or the IO (or to some combination of one or more State(s) or IO(s) or both).[59] If attribution is established, human agents of the State or the IO need to discern whether a breach occurred. This exercise entails assessing the conduct against applicable law. Finally, if the occurrence of a breach is established, human agents of the State or the IO evaluate whether or not the circumstances preclude the wrongfulness of the breach.[60]

Precondition #8: Human Agents of the State or the International Organization Facilitate Incurrence of Responsibility

The eighth precondition concerns situations in which a breach – the wrongfulness of which is not precluded by the circumstances – is established. The precondition is that, where such a breach is established, human agents of the State or the IO arguably need to facilitate incurrence of responsibility of the State or the IO concerning the breach. As part of the process to facilitate such incurrence of responsibility, human agents of the State or the IO may arguably need to impose relevant consequences on the State or the IO. Those consequences may relate, for example, to cessation or reparation (or both) by the State or the IO.[61]

Summary

Suppose that the various premises underlying the above-elaborated preconditions are valid. In that case, the absence of one or more of the following conditions may be preclusive of an element integral to respect for international law by the State or the IO:

[59] For an exploration of certain legal aspects of attribution in relation to 'cyber operations' (which may or may not involve AI-related tools or techniques), see HG Dederer and T Singer, 'Adverse Cyber Operations: Causality, Attribution, Evidence, and Due Diligence' (2019) 95 *ILS* 430, 435–466.

[60] See (D)ARSIWA (n 40) ch V; (D)ARIO (n 30) ch V.

[61] See (D)ARSIWA (n 40), Articles 30–31; (D)ARIO (n 30), Articles 30–31.

1. An exercise and implementation of international law by human agents of the State or the IO in relation to the conduct that forms an employment in an armed conflict of an AI-related tool or technique;
2. A sufficient understanding by human agents of the State or the IO of the technical performance and effects of the employed AI-related tool or technique in relation to the circumstances of use and the socio-technical system through which the tools or techniques are employed;
3. Discernment by human agents of the State or the IO of the law applicable to the State or the IO in relation to the employment;
4. An assessment by human agents of the State or the IO whether the anticipated employment would conform with applicable law in relation to the anticipated specific circumstances and conditions of the employment;
5. Imposition by human agents of the State or the IO of limitations or prohibitions (or both) as required by applicable law in respect of the employment;
6. An assessment by human agents of the State or the IO after employment as to whether or not the employment complied with applicable law;
7. An assessment by human agents of the State or the IO as to whether or not the conduct constitutes a violation, and, if so, (also) an evaluation by human agents of the State or the IO as to whether or not the international legal responsibility of the State or the IO is engaged; or
8. Facilitation by human agents of the State or the IO of the incurrence of responsibility – including imposition of relevant consequences on the State or the IO – where such responsibility is established.

2. *Preconditions Concerning Non-Involved Humans and Entities Related to Respect for International Law by a State or an International Organization*

In this sub-section, I seek to outline some preconditions underlying elements that are arguably necessary for non-involved humans and related entities to (help) ensure respect for international law by a State or an international organization whose conduct forms an employment in an armed conflict of an AI-related tool or technique. Such non-involved people might include, for example, legal advisers from another State or another IO or judges on an international court seized with proceedings instituted by one State against another State.

Precondition #1: Humans Are Legal Agents

As with the previous sub-section, the first precondition here is that humans are arguably the agents for the exercise and implementation of international law applicable to the State or the IO whose conduct forms an employment of an AI-related tool or technique.[62] This precondition is premised on the notion that existing international law presupposes that the functional exercise and implementation of international law to a State or an IO by a human (and by an entity to which that human is connected) not involved in relevant conduct is reserved solely to humans. According to this approach, that primary exercise and implementation of international law may not be partly or wholly reposed in non-human (artificial) agents.

[62] See Switzerland, 'Towards a "Compliance-Based" Approach', above (n 53); US DoD OGC, *Law of War Manual*, above (n 53).

Precondition #2: Humans Discern the Existence of Conduct that Forms an Employment of an AI-Related Tool or Technique

The second precondition is that humans not involved in the conduct of the State or the IO arguably need to discern the existence of the conduct that forms an employment in an armed conflict of an AI-related tool or technique attributable to the State or the IO. To instantiate this precondition, the conduct must be susceptible to being discerned by (non-involved) humans.

Precondition #3: Humans Attribute Relevant Conduct of One or More States or International Organizations to the Relevant Entity or Entities

The third precondition is that humans not involved in the conduct of the State or the IO arguably need to attribute the conduct that forms an employment in an armed conflict of an AI-related tool or technique by or on behalf of the State or the IO to that State or that IO (or to some combination of State(s) or IO(s) or both). To instantiate this precondition, the conduct undertaken by or on behalf of the State or the IO must be susceptible to being attributed by (non-involved) humans to the State or the IO.

Precondition #4: Humans Discern the Law Applicable to Relevant Conduct

The fourth precondition is that humans not involved in the conduct of the State or the IO arguably need to discern the law applicable to the conduct that forms an employment in an armed conflict of an AI-related tool or technique attributable to the State or the IO. To instantiate this precondition, the legal provisions applicable to the State or the IO to which the relevant conduct is attributable must be susceptible to being discerned by (non-involved) humans. For example, where an employment of an AI-related tool or technique by a State occurs in connection with an armed conflict to which the State is a party, humans not involved in the conduct may need to discern whether the State has become party to a particular treaty and, if not, whether a possibly relevant rule reflected in that treaty is otherwise binding on the State, for example through customary international law.

Precondition #5: Humans Assess Potential Violations

The fifth precondition is that humans not involved in the conduct that forms an employment in an armed conflict of an AI-related tool or technique attributable to the State or the IO arguably need to assess possible violations by the State or the IO concerning that conduct.

To make that assessment, (non-involved) humans need to discern, first, whether or not the relevant conduct is attributable to the State or the IO. To instantiate this aspect of the fifth precondition, the conduct forming the employment in an armed conflict of an AI-related tool or technique must be susceptible to being attributed by (non-involved) humans to the State or the IO.

If attribution to the State or the IO is established, (non-involved) humans need to discern the existence or not of the occurrence of a breach. To instantiate this aspect of the fifth precondition, the conduct forming the employment in an armed conflict of an AI-related tool or technique by the State or the IO must be susceptible to being evaluated by (non-involved) humans as to whether or not the conduct constitutes a breach.

If the existence of a breach is established, (non-involved) humans need to assess whether or not the circumstances preclude the wrongfulness of the violation. To instantiate this aspect of the fifth precondition, the conduct forming the employment in an armed conflict of an AI-related tool or technique must be susceptible to being evaluated by (non-involved) humans as to whether or not the specific circumstances preclude the wrongfulness of the breach.

Precondition #6: Humans (and an Entity or Entities) Facilitate Incurrence of Responsibility
The sixth precondition is that humans (and an entity or entities) not involved in the conduct that forms an employment in an armed conflict of an AI-related tool or technique attributable to the State or the IO arguably need to facilitate incurrence of responsibility for a breach the wrongfulness of which is not precluded by the circumstances. In practice, responsibility may be incurred through relatively more formal channels (such as through the institution of State-vs.-State legal proceedings) or less formal modalities (such as through non-public communications between States).

As part of the process to facilitate incurrence of responsibility, (non-involved) humans arguably need to impose relevant consequences on the responsible State or IO. Typically, those humans do so by acting through a legal entity to which they are attached or through which they otherwise (seek to) ensure respect for international law – for example, consider legal advisers of another State, another IO, or judge on an international court. The consequences may relate to (among other things) cessation and reparations.

Regarding cessation, the responsible State or IO is obliged to cease the act, if it is continuing, and to offer appropriate assurances and guarantees of non-repetition, if circumstances so require.[63] To instantiate this aspect of the sixth precondition, the conduct forming the employment in an armed conflict of an AI-related tool or technique must be susceptible to being evaluated by (non-involved) humans as to whether or not the conduct is continuing; furthermore, the conduct must (also) be susceptible to being subject to an offer of appropriate assurances and guarantees of non-repetition, if circumstances so require.

Regarding reparation, the responsible State or IO is obliged to make full reparation for the injury caused by the internationally wrongful act.[64] To instantiate this aspect of the sixth precondition, the conduct forming the employment in an armed conflict of an AI-related tool or technique must be susceptible both to a determination by (non-involved) humans of the injury caused and to the making of full reparations in respect of the injury.

Summary
Suppose that the various premises underlying the above-elaborated preconditions are valid. In that case, the absence of one or more of the following conditions may be preclusive of an element integral to (non-involved) humans and entities helping to ensure respect for international law by a State or an IO where the latter's conduct forms an employment in an armed conflict of an AI-related tool or technique:

1. An exercise and implementation by (non-involved) humans of international law applicable to the State or IO in relation to the conduct;
2. Discernment by (non-involved) humans of the existence of the relevant conduct attributable to the State or the IO;
3. An attribution by (non-involved) humans of the relevant conduct undertaken by or on behalf of the State or the IO;
4. Discernment by (non-involved) humans of the law applicable to the relevant conduct attributable to the State or the IO;
5. An assessment by (non-involved) humans of possible violations committed by the State or the IO in connection with the relevant conduct; or

[63] See (D)ARSIWA (n 40) Article 30; (D)ARIO (n 30) Article 30.
[64] See (D)ARSIWA (n 40) Article 31; (D)ARIO (n 30) Article 31.

6. Facilitation by (non-involved) humans of an incurrence of responsibility of the responsible State or the responsible IO for a breach the wrongfulness of which is not precluded by the circumstances.

3. *Preconditions Concerning Respect for the ICC Statute*

In the above sub-sections, I focused on respect for international law concerning employments in armed conflicts of AI-related tools and techniques by or on behalf of a State or an IO, whether the issue concerns respect for international law by those involved in the conduct (IV 1) or whether it concerns those not involved in the conduct (IV 2). In this sub-section, I seek to outline some preconditions underlying elements that are arguably necessary for respect for the ICC Statute. As noted previously, under the ICC Statute, individual criminal responsibility may arise for certain international crimes, and an employment in an armed conflict of an AI-related tool or technique may constitute, or otherwise contribute to, such a crime. In this section, I use the phrase 'ICC-related human agents' to mean humans who exercise and implement international law in relation to an application of the ICC Statute. Such human agents may include (among others) the court's prosecutors, defense counsel, registrar, and judges.

Precondition #1: Humans Are Legal Agents

The first precondition is that humans are arguably the agents for the exercise and implementation of international law applicable in relation to international crimes – including under the ICC Statute – arising from conduct that forms an employment in an armed conflict of an AI-related tool or technique.[65] (Of the four categories of crimes under the ICC Statute, strictly speaking only war crimes by definition must necessarily be committed in connection with an armed conflict. Nonetheless, the other three categories of crimes under the ICC Statute may be committed in connection with an armed conflict.) This precondition is premised on the notion that existing international law presupposes that the functional exercise and implementation of international law to the conduct of a natural person is reserved solely to humans (and, through them, to the entity or entities, such as an international criminal tribunal, to which those humans are attached). According to this approach, this primary exercise and implementation of international law may not be partly or wholly reposed in non-human (artificial) agents.

Precondition #2: Humans Discern the Existence of Potentially Relevant Conduct

The second precondition is that ICC-related human agents arguably need to discern the existence of conduct that forms an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person. For this precondition to be instantiated, such conduct must be susceptible to being discerned by relevant ICC-related human agents.

Precondition #3: Humans Determine Whether the ICC May Exercise Jurisdiction

The third precondition is that ICC-related human agents arguably need to determine whether or not the court may exercise jurisdiction in relation to an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person. The court may exercise jurisdiction only over natural persons.[66] Furthermore, the ICC may exercise jurisdiction only where the

[65] See Switzerland, 'Towards a "Compliance-Based" Approach', above (n 53); US DoD OGC, *Law of War Manual*, above (n 53).
[66] ICC Statute, Article 25(1).

relevant elements of jurisdiction are satisfied.[67] To instantiate the third precondition, conduct that forms an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person must be susceptible to being evaluated by relevant ICC-related human agents as to whether or not the conduct is attributable to one or more natural persons over whom the court may exercise jurisdiction.

Precondition #4: Humans Adjudicate Individual Criminal Responsibility

The fourth precondition is that ICC-related human agents arguably need to adjudicate whether or not an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person subject to the jurisdiction of the court constitutes, or otherwise contributes to, an international crime over which the court has jurisdiction. For the fourth precondition to be instantiated, such conduct must be susceptible to being evaluated by relevant ICC-related human agents – in pre-trial proceedings, trial proceedings, and appeals-and-revision proceedings – as to whether or not (among other things) the conduct satisfies the 'material'[68] and 'mental'[69] elements of one or more crimes and whether the conduct was undertaken through a recognized mode of responsibility.[70]

Precondition #5: Humans Facilitate the Incurrence of Individual Criminal Responsibility

The fifth precondition is that ICC-related human agents arguably need to facilitate incurrence of individual criminal responsibility for an international crime where such responsibility is established. As part of the process to facilitate the incurrence of such responsibility, relevant ICC-related humans need to (among other things) facilitate the imposition of penalties on the responsible natural person(s).[71] For the fifth precondition to be instantiated, the conduct underlying the establishment of individual criminal responsibility needs to be susceptible to being subject to the imposition of penalties on the responsible natural person(s).

Summary

Suppose that the various premises underlying the above-elaborated preconditions are valid. In that case, the absence of one or more of the following conditions – in relation to an employment in an armed conflict of an AI-related tool or technique that constitutes, or otherwise contributes to, an international crime – may be preclusive of respect for the ICC Statute:

1. An exercise and implementation of international law by one or more relevant ICC-related human agents concerning the conduct;
2. Discernment by one or more relevant ICC-related human agents of the conduct that forms an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person;
3. A determination by one or more relevant ICC-related human agents whether or not the court may exercise jurisdiction in respect of an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person;
4. An adjudication by relevant ICC-related human agents whether or not an employment in an armed conflict of an AI-related tool or technique ascribable to a natural person subject

[67] See ICC Statute, Articles 5–19.
[68] See ICC Statute, Articles 6–8 *bis*.
[69] See ICC Statute, Article 30.
[70] See ICC Statute, Articles 25, 28.
[71] ICC Statute, Article 77.

to the jurisdiction of the court constitutes, or otherwise contributes to, an international crime over which the court has jurisdiction; or

5. Facilitation by one or more relevant ICC-related human agents of an incurrence of individual criminal responsibility – including the imposition of applicable penalties on the responsible natural person(s) – where such responsibility is established.

V. CONCLUSION

An employment in an armed conflict of an AI-related tool or technique that is attributable to a State, an IO, or a natural person (or some combination thereof) is governed at least in part by international law. It is well established that international law sets out standard assumptions of responsibility for the conduct of States and IOs. It is also well established that it is on the basis of those assumptions that specific legal provisions exist and are applied in respect of those entities. International law also arguably sets out particular standard assumptions of criminal responsibility for the conduct of natural persons. It may be contended that it is on the basis of those assumptions that the ICC Statute exists and is applied.

Concerning the use of AI in armed conflicts, at least three categories of human agents may be involved in seeking to ensure that States, IOs, or natural persons respect applicable law. Those categories are the human agents acting on behalf of the State or the IO engaging in relevant conduct; human agents not involved in such conduct but who nevertheless (seek to) ensure respect for international law in relation to that conduct; and human agents who (seek to) ensure respect for the ICC Statute. Each of those human agents may seek to respect or ensure respect for international law in connection with a legal entity to which they are attached or through which they otherwise act.

'Responsible AI' is not a term of art in international law, at least not yet. It may be argued the preconditions arguably necessary to respect international law – principally in the sense of applying and observing international law and facilitating incurrence of responsibility for violations – ought to be taken into account in formulating notions of 'responsible AI' pertaining to relevant conduct connected with armed conflict. Regarding those preconditions, it may be argued that, under existing law, humans are the (at least primary) legal agents for the exercise and implementation of international law applicable to an armed conflict. It may also be submitted that, under existing law, an employment in an armed conflict of an AI-related tool or technique needs to be susceptible to being (among other things) administered, discerned, attributed, understood, and assessed by one or more human agent(s).[72]

Whether – and, if so, the extent to which – international actors will commit in practice to instantiating the preconditions arguably necessary for respecting international law pertaining to an employment in an armed conflict of an AI-related tool or technique will depend on factors that I have not expressly addressed in this chapter but that warrant extensive consideration.

[72] See DA Lewis, 'Three Pathways to Secure Greater Respect for International Law Concerning War Algorithms' (Harvard Law School Program on International Law and Armed Conflict, December 2020) https://dash.harvard.edu/bitstream/handle/1/37367712/Three-Pathways-to-Secure-Greater-Respect.pdf?sequence=1&isAllowed=y; V Boulanin, L Bruun, and N Goussac, 'Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction' (Stockholm International Peace Research Institute, 2021) www.sipri.org/sites/default/files/2021-06/2106_aws_and_ihl_0.pdf.