6

Human Psychology and Robot Evidence in the Courtroom, Alternative Dispute Resolution, and Agency Proceedings

SARA SUN BEALE AND HAYLEY LAWRENCE*

I Introduction

In the courtroom, the phrases artificial intelligence (AI) and robot witnesses ("robo-witnesses") conjure up images of a Star Wars-like, futuristic world with autonomous robots like C3PO taking the witness stand. Although testimony from a robo-witness may be possible in the distant future, many other kinds of evidence produced by AI are already becoming more common.

Given the wide and rapidly expanding range of activities being undertaken by robots, it is inevitable that robot-generated evidence and evidence from human witnesses who interacted with or observed robots will be presented in legal forums. This chapter explores the effects of human psychology on human–robot interactions (HRIs) in legal proceedings. In Section II, we review the research on HRI in other contexts, such as market research and consumer interactions. In Section III, we consider the effect the psychological responses detailed in Section II may have in litigation.

We argue that human responses to robot-generated evidence will present unique challenges to the accuracy of litigation, as well as ancillary goals such as fairness and transparency, but HRI may also enhance accuracy in other respects. For our purposes, the most important feature of HRI is the human tendency to anthropomorphize robots. Anthropomorphization can generate misleading impressions, e.g., that robots have human-like emotions and motives, and this tendency toward anthropomorphization can be manipulated by designing robots to make them appear more

* Sara Sun Beale, Charles L. B. Lowndes Professor of Law, Duke Law School; Hayley N. Lawrence, JD, LLM, Duke Law School, 2021.

trustworthy and believable. The degree of distortion caused by anthropomorphization will vary, depending on the design of the robot and other situational factors, like how the interaction is framed. The effects of anthropomorphization may be amplified by the simulation heuristic, i.e., how people estimate the likelihood that something happened based on how easy it is for them to imagine it happening, and the psychological preference for direct evidence over circumstantial evidence.[1] Moreover, additional cognitive biases may distort fact-finding or attributions of liability when humans interact with or observe robots.

On the other hand, robot-generated evidence may offer unique advantages if it can be presented as direct evidence via a robo-witness, because of the nature of a robo-witness's memory compared to that of a human eyewitness. We have concerns, however, about the degree to which the traditional methods of testing the accuracy of evidence, particularly cross-examination, will be effective for robot-generated evidence. It is unclear whether lay fact-finders, who are prone to anthropomorphize robots, will be able to understand and evaluate the information generated by complex algorithms, particularly those using unsupervised learning models.

Although it has played a limited role in litigation, AI evidence has been used in other legal forums. Section IV compares the use of testimony from autonomous vehicles (AVs) in litigation with the use of similar evidence in alternative dispute resolution (ADR) and the National Transportation Safety Board (NTSB). These contrasting legal infrastructures present an opportunity to examine AI evidence through a different lens. After comparing and contrasting AI testimony in ADR and NTSB proceedings with traditional litigation, the chapter suggests that the presence of expert decision-makers might help mitigate some of the problems with HRI, although other aspects of the procedures in each forum still raise concerns.

II The Psychology of HRI in Litigation

Although there is no universally agreed-upon definition of "robot," for our purposes, a robot is "an engineered machine that senses, thinks, and acts."[2] Practically speaking, that means the robot must "have sensors, processing ability that emulates some aspect of cognition," and the capacity

[1] See Section III.D.4.

[2] Patrick Lin, Keith Abney, & George Bekey, "Robot Ethics: Mapping the Issues for a Mechanized World" (2011) 175:5–6 *Artificial Intelligence* 942 at 943.

to act on its decision-making.[3] A robot must be equipped with programming that allows it to independently make intelligent choices or perform tasks based on environmental stimuli, rather than merely following the directions of a human operator, like a remote controlled car.[4] Under our definition, robots need not be embodied, i.e., they need not occupy physical space or have a physical presence. Of course, the fictitious examples of R2D2 and C3P0 fit our definition, but so too do the self-driving, guided steering, or automatic braking features in modern cars.

II.A Anthropomorphism

The aspect of HRI with the greatest potential to affect litigation is the human tendency to anthropomorphize robots.[5] Despite knowing that robots do not share human consciousness, people nevertheless tend to view robots as inherently social actors. As a result, people often unconsciously apply social rules and expectations to robots, assigning to them human emotions and sentience.[6] People even apply stereotypes and social heuristics to robots[7] and use the same language to describe interactions with robots and humans.[8] This process is unconscious and instantaneous.[9]

Rather than operating like an on-off switch, there are degrees of anthropomorphization, and the extent to which people anthropomorphize depends on several factors, including framing, interactivity or animacy,

[3] Ibid.

[4] Ibid.

[5] Kate Darling, "'Who's Johnny?': Anthropomorphic Framing in Human–Robot Interaction, Integration, and Policy" in Patrick Lin, Keith Abney, & Ryan Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (New York, NY: Oxford University Press, 2017) 173 ["Who's Johnny"] at 173; see Chapter 13 in this volume.

[6] Ibid.

[7] Aaron Powers & Sara Keisler, "The Advisor Robot: Tracing People's Mental Model from a Robot's Physical Attributes" (paper delivered at the International Conference on Human–Robot Interaction, March 2–3, 2006), HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction (New York, NY: Association for Computing Machinery, 2006) 218, www.cs.cmu.edu/~kiesler/publications/2006pdfs/2006_advisor-robot.pdf ["Advisor Robot"].

[8] Susan Fussell, Sara Kiesler, Leslie D. Setlock *et al.*, "How People Anthropomorphize Robots" (paper delivered at the International Human–Robot Interaction Conference, March 12–15, 2008), HRI '08: Proceedings of the 3rd ACM/IEEE International Conference on Human–Robot Interaction (New York, NY: Association for Computing Machinery, 2008) 145 at 149, www.cs.cmu.edu/~./kiesler/publications/2008pdfs/2008_anthropomorphize-bots.pdf.

[9] "Advisor Robot", note 7 above, at 2.

physical embodiment and presence, and appearance. Furthermore, these factors interact with one another. The presence (or absence) of a given characteristic impacts the anthropomorphizing effect of the other present characteristics.

II.A.1 Framing

How an HRI is framed significantly impacts human responses and perceptions about the robot and the interaction itself. Framing therefore has the potential to interfere with the accuracy of the litigation process when robot-generated evidence is presented. Framing generally refers to the way a human observer is introduced to an interaction, and in the case of robot-generated evidence, to a robot before the interaction actually begins. For example, does the robot have a name? Is the name endearing or human-like, e.g., "Marty" versus "Model X"? Is the robot assigned a gender? Is the robot given a backstory? What job or role is the robot intended to fulfil? Framing immediately impacts the human's perception of a robot. Humans use that introductory information to form a mental model about a robot, much as they do for people, assigning to it stereotypes, personal experiences, and human emotions through anthropomorphization.[10]

Two experiments demonstrate the power of framing to establish trust and create emotional attachments to robots. The first experiment involved participants riding in AVs, which are robots by our definition, and it demonstrates how framing can impact people's trust in a robot and how much blame they assign to it.[11] Each test group was exposed to a simulated crash that was unavoidable and clearly caused by another simulated driver. Prior to the incident, participants who had received anthropomorphic framing information about the car, including a name, a gendered identity, and a voice through human audio files, trusted the car more than participants who had ridden in a car with identical driving capabilities but for which no similar framing information had been provided ("agentic condition") and more than those in the "normal" condition who operated the car themselves, i.e., no autonomous capabilities.[12] After the incident, participants reported that they trusted the

[10] "Who's Johnny", note 5 above, at 180.

[11] Adam Waytz, Joy Heafner, & Nicholas Epley, "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle" (2014) 52 *Journal of Experimental Social Psychology* 113 at 115.

[12] Ibid.

anthropomorphically framed car more even though the only difference between the two conditions was the car having humanized qualities. Subjects in the anthropomorphized group also blamed the vehicle for the incident significantly less than the agentic group, perhaps because they unconsciously perceived the car as more thoughtful. Conversely, subjects in the normal condition who operated the car themselves assigned very little blame to the car. This makes sense because "[a]n object with no agency cannot be held responsible for any actions."[13] It is important that the anthropomorphized condition group perceived the car as more thoughtful, which mitigated some of the responsibility imputed to the vehicle.

The second experiment demonstrates that the way a robot's relationship to humans is framed, even by something as simple as giving the robot a name, can seriously impact the level of emotional attachment humans feel toward it. Participants were asked to observe a bug-like robot and then to strike it with a mallet.[14] The robot was introduced to one group of study participants with a name and an appealing backstory. "This is Frank. Frank is really friendly, but he gets distracted easily. He's lived at the Lab for a few months now. His favorite color is red."[15] The participants who experienced this anthropomorphic framing demonstrated higher levels of empathy and concern for the robot, showing emotional distress and a reluctance to hit it.[16]

Additionally, framing may impact whether, and to what degree, humans assume a robot has agency or free will. Anthropomorphism drives humans to impute at least a basic level of human "free will" to robots.[17] In other words, people assume that a robot makes at least some of its choices independently rather than as a simple result of its internal programming. This understanding is, of course, flawed. Although AI "neural networks" are modeled after the human brain to identify patterns and make decisions,

[13] Ibid.

[14] "Who's Johnny", note 5 above, at 181.

[15] Kate Darling, Palash Nandy, & Cynthia Breazeal, "Empathic Concern and the Effect of Stories in Human–Robot Interaction" (paper delivered at the IEEE International Workshop on Robot and Human Communication (RO-MAN), August 31–September 1, 2015), 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (Kobe, Japan: IEEE, 2015) 770 at 3, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2639689.

[16] Ibid. at 11–12.

[17] Neil Richards & William Smart, "How Should the Law Think about Robots?" in Ryan Calo, A. Michael Froomkin, & Ian Kerr (eds.), *Robot Law* (Cheltenham, UK: Edward Elgar, 2016) [*Robot Law*] 3 at 18.

robots do not consciously think and make choices as we do.[18] As robots operate more autonomously and are equipped with more anthropomorphous characteristics, humans will likely perceive them as having more agency or free will.[19]

II.A.2 Interactivity or Animacy

The interactivity or animacy of a robot also has a significant effect on HRI. Anthropomorphization drives people to seek social connections with robots,[20] and our innate need for social connection also causes humans to infer from a robot's verbal and non-verbal "expressions" that it has "emotions, preferences, motivations, and personality."[21] Social robots can now simulate sound, movement, and social cues that people automatically and subconsciously associate with human intention and states of mind.[22] Robots can motivate people by mimicking human emotions like anger, happiness, or fear, and demonstrate a pseudo-empathy by acting supportively.[23] They can apply peer pressure or shame humans into doing or not doing something.[24]

Humans form opinions about others based on voice and speech patterns,[25] and the same responses, coupled with anthropomorphization, can be used to make judgments about robots' speech. Many robots

[18] Instead, neural networks are comprised of a series of complex decision trees that are programmed to react according to environmental stimuli. Larry Hardesty, "Explained: Neural Networks," *MIT News* (April 14, 2017), http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.

[19] Matthias Scheutz, "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots" in Patrick Lin, Keith Abney, & George Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (London, UK: MIT Press, 2012) 205 at 211–214.

[20] Ibid. at 205–221.

[21] Serena Marchesi, Davide De Tommaso, Jairo Perez-Osorio *et al.*, "Belief in Sharing the Same Phenomenological Experience Increases the Likelihood of Adopting the Intentional Stance Toward a Humanoid Robot" (2022) 3:3 *Technology, Mind, and Behavior* 1 (finding subjects with exposure to a human-like robot were more likely to rate the robot's actions as intentional).

[22] "Who's Johnny", note 5 above, at 175–176.

[23] Brian Jeffrey Fogg, "Computers as Persuasive Social Actors" in Brian Jeffrey Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do* (San Francisco, CA: Morgan Kaufmann Publishers, 2003) 89 ["Persuasive Social Actors"] at 100.

[24] Ibid.

[25] Phil McAleer, Alexander Todorov, & Pascal Belint, "How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices" (2014) 9:3 *PLoS ONE* 1; see also "Advisor Robot", note 7 above, at 1.

now communicate verbally, using verbal communication to persuade humans, establish a "relationship," or convey moods or a personality.[26] Certain styles of speech, accents, and vernacular are perceived as more authoritative, trustworthy, persuasive, or intelligent.[27]

II.A.3 Physical Presence and Physical Embodiment

Physical presence and physical embodiment also impact the extent to which people anthropomorphize a robot. A physically present robot is one that shares the same physical space with you. A physically embodied robot is one that has some sort of physical manifestation. A robot may be physically embodied, but not physically present. A familiar example is the Roomba vacuum robot. A Roomba in your house is physically present and physically embodied. But if you interact with C3P0, the gold robot from Star Wars, via video conference, C3P0 is physically embodied, but not physically present. Instead, he is telepresent. Lastly, Apple's Siri is an example of a robot that is neither physically present nor physically embodied. The Siri virtual assistant is a voice with neither a physical appearance nor an embodiment outside the iPhone.

In experimental settings, a physically present, embodied robot affected HRI more than its non-embodied or non-present counterparts.[28] The combination of the robot's presence and embodiment fostered favorable attitudes among study participants. These findings are consistent with the assumptions that people perceive robot agents as social actors and typically prefer face-to-face interactions.[29] A review of multiple studies found that participants had more favorable attitudes toward co-present, physically embodied robots than toward telepresent robots, and that physically embodied robots were more persuasive and more trustworthy than

[26] "Persuasive Social Actors", note 23 above, at 101.

[27] See generally, Andrea Morales, Maura Scott, & Eric Yorkston, "The Role of Accent Standardness in Message Preference and Recall" (2012) 41:1 *Journal of Advertising* 33 ["Accent Standardness"] at 34 (studying people's accent preferences, noting, e.g., that "[s]ociolinguistic research shows that speakers with standard English accents are seen as having high social status and as being competent, smart, educated, and formal").

[28] Jamy Li, "The Benefit of Being Physically Present: A Survey of Experimental Works Comparing Copresent Robots, Telepresent Robots, and Virtual Agents" (2015) 77 *International Journal of Human-Computer Studies* 23 ["The Benefit"] at 33.

[29] "Accent Standardness", note 27 above, at 34.

their telepresent counterparts.[30] There was, however, no statistically significant difference between human perception of telepresent robots and non-embodied virtual agents like Siri. Overall, participants favored the co-present robot to the virtual agent and found the co-present robot more persuasive, even when its behavior was identical to that of the virtual agent. People paid more attention to the co-present robot and were more engaged in the interaction.

II.A.4 Appearance

Because of the power of anthropomorphism, the appearance or features of an embodied robot can influence whether it is viewed as likeable, trustworthy, and persuasive.

II.A.4.a Robot Faces Whether a robot is given a face, and what that face looks like, will have a significant impact on HRI. Humans form impressions almost instantly, deciding whether a person is attractive and trustworthy within one-tenth of a second of seeing their face.[31] Because humans incorrectly assume that robots are inherently social creatures, we make judgments about robots based on their physical attributes using many of the same mental shortcuts that we use for humans. Within the first two minutes of a human–robot interaction or observation, "people create a coherent, plausible mental model of the robot," based primarily on its physical appearance and interactive features like voice.[32]

Because humans derive many social cues from facial expressions, a robot's head and face are the physical features that most significantly affect HRI.[33] People notice the same features in a robot face that they notice about a human one: eye color and shape, nose size, etc.,[34] and researchers already have a basic understanding of what esthetic features humans like

[30] Twenty-four out of twenty-nine studies surveyed confirmed this point: see "The Benefit", note 28 above, at 33.

[31] Chad Boutin, "Snap Judgments Decide a Face's Character, Psychologist Finds," *Princeton University* (August 22, 2006), www.princeton.edu/news/2006/08/22/snap-judgments-decide-faces-character-psychologist-finds.

[32] See "Advisor Robot", note 7 above, at 6.

[33] Julia Fink, "Anthropomorphism and Human Likeness in the Design of Robots and Human–Robot Interaction" (paper delivered at the 4th International Conference, ICSR 2012, October 29–31, 2012) in Shuzi Sam Ge, Oussama Khatib, John-John Cabibihan *et al.* (eds.), *Social Robotics* (Berlin, Germany: Springer, 2012) 199 at 203 (noting that "most non-verbal cues are mediated through the face").

[34] People notice the same features they would notice unconsciously about a human face when they view a robot's face. Carl DiSalvo, Francine Gemperle, Jodi Forlizzi *et al.*, "All Robots

or dislike in robots. For example, robots with big eyes and "baby faces" are perceived as naïve, honest, kind, unthreatening, and warm.[35] Researchers are also studying how features make robot heads and faces more or less likeable and persuasive.[36] Manipulating the relative size of the features on a robot's head had a significant effect on not only study participants' evaluation of a robot, but also on whether they trusted it and would be likely to follow its advice.[37] A robot with big eyes was perceived as warmer and more honest and participants were thus more likely to follow its health advice.

II.A.4.b Physical Embodiment and Interactive Style When interacting with physically embodied robots, human subjects report that interactions with responsive robots, those with animated facial expressions, social gaze, and/or mannerisms, feel more natural and enjoyable than interactions with unanimated robots.[38] Embodied robots with faces can be programed to directly mirror subjects' expressions, or to indirectly mirror these expressions based on the robot's evaluation of the subject's perceived security, arousal, and autonomy. Study participants rated indirect mirroring robots highest for empathy, trust, sociability, and enjoyment,[39] and rated indirect mirroring and mirroring robots higher than the non-mirroring robots in empathy, trust, sociability, enjoyment, anthropomorphism, likeability, and intelligence.[40]

Generally, lifelike physical movement of robots, including "social gaze," or when a robot's eyes follow the subject it's interacting with,[41]

Are Not Created Equal: The Design and Perception of Humanoid Robot Heads" (paper delivered at the Conference on Designing Interactive Systems, June 25–28, 2002), DIS '02: Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (New York, NY: Association for Computing Machinery, 2002) 321 at 322, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.7443&rep=rep1&type=pdf ["Not Created Equal"].

35 "Advisor Robot", note 7 above, at 6.

36 "Persuasive Social Actors", note 23 above, at 92–93.

37 "Advisor Robot", note 7 above, at 6.

38 For a study examining the correlation between a co-present robot's emotional nonverbal response and a human's anthropomorphic response, see Friederike Eyssel, Frank Hegel, Gernot Horstmann *et al.*, "Anthropomorphic Inferences from Emotional Nonverbal Cues: A Case Study" (paper delivered at the 19th International Conference, September 13–15, 2010), 19th International Symposium in Robot and Human Interactive Communication (Viareggio, Italy: IEEE, 2010) 646 at 646.

39 "Not Created Equal", note 34 above, at 353–354 and 356.

40 Ibid.

41 See Debora Zanatto, Massimiliano Patacchiola, Jeremy Goslin *et al.*, "Priming Anthropomorphism: Can the Credibility of Humanlike Robots Be Transferred to

gestures, and human-like facial expressions, are highly correlated with anthropomorphic projection.[42] When those movements closely match humans' non-verbal cues, humans perceive robots as more human-like. This matching behavior, exemplified through non-verbal cues, like facial expressions, gestures, e.g., nodding, and posture, is known as behavioral mimicry.[43] Behavioral mimicry is critical for establishing rapport and empathy in human interactions,[44] and this phenomenon extends to HRI as well.[45]

II.B Other Cognitive Biases

A variety of other cognitive errors may distort fact-finding or the imposition of liability for the conduct of robots. For example, in experimental settings, subjects tended to blame human actors more than robots for the same conduct.

One study tested the allocation of blame for a hypothetical automobile accident in which a pedestrian has been killed by an automated car, and both the human driver and the automated system, a robot for our purposes, have made errors.[46] The "central finding is that in cases where a human and a machine share control of the car in hypothetical situations, less blame is attributed to the machine when both drivers make errors."[47]

Non-Humanlike Robots?" (paper delivered at the 2016 11th ACM/IEEE Conference on HRI, March 7–10, 2016), 2016 11th ACM/IEEE International Conference on Human–Robot Interaction (Christchurch: IEEE, 2016) 543 at 543–544 (finding that people perceived an anthropomorphic robot as more credible than its non-anthropomorphic counterpart when it used social gaze, as measured by willingness to change their response to a question based on information provided by the robot).

[42] "Who's Johnny", note 5 above, at 174, 175–176.

[43] Elise Owens, Ferguson W. H. McPharlin, Nathan Brooks *et al.*, "The Effects of Empathy, Emotional Intelligence and Psychopathy on Interpersonal Interactions" (2018) 25:1 *Psychiatry, Psychology and Law* 1 at 1–2.

[44] Ibid.

[45] Barbara Gonsior, Stefan Sosnowski, Christoph Mayer *et al.*, "Improving Aspects of Empathy and Subjective Performance for HRI through Mirroring Facial Expressions" (paper delivered at IEEE RO-MAN Conference, July 31–August 3, 2011), 2011 RO-MAN (Atlanta, GA: IEEE, 2011) 350 at 351, www.researchgate.net/publication/224256284_Improving_aspects_of_empathy_and_subjective_performance_for_HRI_through_mirroring_facial_expressions.

[46] Edmond Awad, Sydney Levine, Max Kleiman-Weiner *et al.*, "Drivers Are Blamed More than Their Automated Cars When Both Make Mistakes" (2020) 4:2 *Nature Human Behaviour* 134 ["Drivers Are Blamed"].

[47] Ibid. at 138.

In all scenarios, subjects attributed less blame to the automatic system when there was a human involved.

Other studies found that in experimental conditions subjects valued algorithmic predictions differently from human input. Coining the term "algorithmic appreciation," the authors of one study found that lay subjects adhered more to advice when they believed it came from an algorithm rather than a person.[48] But this "appreciation" for the algorithm's conclusions decreased when people chose between an algorithm's estimate and their own.[49] Moreover, experienced professionals who made forecasts on a regular basis relied less on algorithmic advice than did lay people, decreasing the professionals' accuracy. But other studies found "algorithmic aversion," with subjects showing more quickly losing confidence in algorithmic than human forecasters, after seeing both make the same mistake.[50]

III The Impact of the Psychology of HRI in Litigation

In this section, we assume that the psychological phenomena described above will occur outside the laboratory setting and, more specifically, in the courtroom. This is a significant assumption because it is difficult to perfectly extrapolate real-world behavior from experimental studies.[51]

The cognitive errors associated with people's tendency to anthropomorphize robots could distort the accuracy and fairness of the litigation process in multiple ways. The current prevalence of these errors may lead to the conclusion that the distortions arising from robot-generated evidence are no greater than those arising from other forms of evidence. Indeed, in some respects, robot-generated evidence might contribute to accuracy because it would be less subject to certain cognitive errors. There remain, however, difficult questions about how well the tools traditionally used

[48] Jennifer Logg, Julia Minson, & Don Moore, "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment" (2019) 151 *Organisational Behavior and Human Decision Processes* 90 ["Algorithm Appreciation"].

[49] Ibid.

[50] Berkeley Dietvorst, Joseph Simmons, & Cade Massey, "Algorithmic Aversion: People Erroneously Avoid Algorithms after Seeing Them Err" (2015) 144:1 *Journal of Experimental Psychology: General* 114.

[51] Cf. "Adversarial Collaboration: An EDGE Lecture by Daniel Kahneman," *EDGE* (February 24, 2022), www.edge.org/adversarial-collaboration-daniel-kahneman (noting difficulty of replicating results of priming experiments).

to test accuracy in litigation can be adapted to robot-generated evidence, as well as questions about the distributional consequences of developing more persuasive robots.

III.A The Impact of Framing and Interactivity

Anthropomorphic framing and tailoring robots to preferences for certain attributes such as speech and voice patterns could distort and impair the accuracy of fact-finding in litigation. Anthropomorphic framing and design can cause humans to develop a false sense of trust and emotional attachment to a robot and may cause fact-finders to incorrectly attribute free will to it. These psychological responses could distort liability determinations if, e.g., jurors who anthropomorphized a robot held it, rather than its designers, responsible for its actions.[52] Indeed, in the automated car study discussed above,[53] because participants perceived the anthropomorphic car as being more thoughtful, they blamed it less than another car with the same automated driving capabilities. Anthropomorphism could also lead fact-finders to attribute moral blame to a robot. For example, in a study in which a robot incorrectly withheld a $20 reward from participants, nearly two-thirds of those participants attributed moral culpability to the robot.[54] Finally, tailoring voice and speech patterns to jurors' preferences could improve a robo-witness's believability, though these features would have no bearing on the reliability of the information provided.

On the other hand, the issues raised by anthropomorphization can be analogized to those already present in litigation. Fact-finders now use heuristics, or mental shortcuts, to evaluate a human witness based on her features, e.g., name, appearance, race, gender, mannerisms. In turn, this information allows jurors to form rapid and often unconscious impressions about the witness's motivations, personality, intelligence, trustworthiness, and believability. Those snap judgments may be equally as unfounded as those a person would make about a robot based on its appearance and framing. And just as a robot's programmed speech patterns may impact the fact-finder's perception of its trustworthiness

[52] See *Robot Law*, note 17 above, at 19.

[53] See notes 46–47 above and accompanying text.

[54] Peter Kahn Jr., Takayuki Kanda, Hiroshi Ishiguro *et al.*, "Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?" (paper delivered at the 7th ACM/IEEE International Conference, March 5–8, 2012), HRI '12: Proceedings of the 7th Annual ACM/IEEE International Conference on Human–Robot Interaction (New York, NY: Association for Computing Machinery, 2012) 33.

and believability, lay or expert human witnesses may be selected or coached to do the same thing. So, although robot-generated evidence and robo-witnesses may differ from their human counterparts, the issues their design and framing present in the litigation context are not entirely novel.

III.B The Impact of Robot Embodiment, Interactivity, and Appearance

Whether a robot is embodied and the form in which it is embodied have a significant impact on human perception. Assuming that these psychological responses extend to the litigation context, it may seem obvious that this would introduce serious distortions into the fact-finding process. But again, this problem is not unique to robots. As noted, humans apply the same unconscious heuristics to human faces, reacting more favorably depending on physical criteria, such as facial proportions, that have no necessary relationship to a witness's truthfulness or reliability. Arguably, the same random distortions could occur for human or robot witnesses. Indeed, assuming equal access to this technology, perhaps the fact that all robot witnesses can be designed to generate positive reactions could eliminate factors that currently distort the fact-finding process in litigation. For example, jurors will not discount the evidence of certain robo-witnesses on grounds such as implicit racial bias, or biases against witnesses who are not physically attractive or well spoken.

III.C The Impact of Other Cognitive Biases

In litigation, other cognitive biases about robots or their algorithmic programming may affect either the attribution of fault or the assessment of the credibility of robot-generated evidence, particularly evidence that is generated by algorithms.

The study discussed earlier, which found a greater tendency to attribute fault to a human rather than an automated system, has clear implications for liability disputes involving automated vehicles. As the authors of the study noted, the convergence of their experimental results with "real world public reaction" to accidents involving automated vehicles suggests that their research findings would have external validity, and that "juries will be biased to absolve the car manufacturer of blame in dual error cases."[55]

[55] "Drivers Are Blamed", note 46 above, at 139–140 (discussing the incidents with Tesla and Uber automated cars).

One of the experiments finding "algorithmic appreciation," which we characterize as the potential for overweighting algorithmic analysis, likely has some direct correlation in litigation, where an algorithm may be seen as more reliable than a variety of human estimates.[56]

III.D Testing the Fidelity of Robot-Generated Evidence in Litigation

Robot-generated evidence already plays a role in litigation proceedings. But how will that dynamic change as robots' capabilities mature to the point of testifying for themselves? We explore the possibilities below.

III.D.1 Impediments to Cross-Examination

It is unclear how adaptable the techniques traditionally used to test a human witness's veracity and reliability are to robot-generated evidence. In particular, the current litigation system relies heavily on cross-examination, based on the assumption that it allows the fact-finder to assess a witness's motivations, behavior, and conclusions. Cross-examination assumes that a witness has motivations, morality, and free will. But robots possess none of those, though fact-finders may erroneously assume that they do. Thus, it may be impossible to employ cross-examination to evaluate the veracity and accuracy of a robo-witness's testimony. Additionally, robot-generated evidence presents two distinct issues: the data itself, and the systems that create the data. Both need to be interrogated, which will require new procedures adapted to the kind of machine or robot evidence in question.[57]

III.D.2 The Difficulty in Evaluating and Challenging Algorithms

Adversarial litigation may also be inadequate to assess defects in a robot's programming, including the accuracy or bias of the algorithm.[58] The quality and accuracy of an algorithm depends on the training instructions and quality of the training data. Designers may unintentionally introduce bias into the algorithm, creating skewed results. For example, algorithms

[56] See "Algorithm Appreciation", note 48 above, at 151.
[57] See generally, Andrea Roth, "Machine Testimony" (2017) 126:1 *Yale Law Journal* 1972; see Chapters 7 and 9 in this volume.
[58] Regarding programmer liability, see Chapter 2 in this volume.

can entrench existing gender biases,[59] and facial recognition software has been criticized for racial biases that severely reduce its accuracy.[60]

It can be extraordinarily difficult to fully understand how an algorithm works, particularly an unsupervised one, in order to verify its accuracy. Unlike supervised learning algorithms, an unsupervised learning algorithm trains on an unlabeled dataset and continuously updates its own training based on environmental stimuli, generally without any external alterations or verification.[61] Although its original code remains the same, the way an unsupervised learning algorithm treats input data may change based on this continuous training. Data goes in and results come out, but how the algorithm reached that result may remain a mystery. Sometimes even the people who originally programmed these algorithms do not fully understand how they operate.

Juries may struggle to understand other complex technology, even with the assistance of experts, and unsupervised learning methods introduce a novel problem into the litigation process because even their creators may not know exactly how they work. This critical gap can only compound the difficulties introduced by anthropomorphism. Experts, even an algorithm's creators, may not be able to understand, let alone explain, how it reached certain conclusions, making it nearly impossible to verify those conclusions in legal proceedings using existing methods.[62]

III.D.3 The Advantages of Robot Memory

Although anthropomorphism can cause distortions, robot-generated evidence is not subject to other cognitive biases that currently impair fact-finding.[63]

The most significant impediment to an accurate evaluation of testimony is pervasive misunderstandings of how memories are formed and recalled. As a foundational matter, many people erroneously assume that our memories operate like recording devices, capturing all the details of a given event, etched permanently in some internal hard drive,

[59] See e.g. Nicol Turner Lee, Paul Resnick, & Genie Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," *Brookings Institution* (May 22, 2019), www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[60] Ibid.

[61] An unsupervised algorithm "tries to make sense by extracting features and patterns on its own."

[62] See Chapter 8 in this volume.

[63] Ibid.

available for instant recall at any moment.[64] But human memory formation is far more complex and fallible. Initially, our memories capture only a very small percentage of the stimuli in our sensory environment.[65] Because there are gaps, we often consciously or subconsciously look for filler information to complete the memory of a given event. Unlike a recording device, which would create a static memory, human memory is dynamic and reconstructive, meaning that post-event interactions or information may alter one's recollection of an event.[66] This susceptibility to influence is called suggestibility.[67] Outside influences can disturb the stability and accuracy of eyewitness memory over time, causing witnesses to misremember details about events they witnessed.[68] Moreover, when people are engaged in memory recall, their recollections are highly suggestible, increasing the likelihood that outside influences will taint their memories.[69]

Although the reliability of human memory depends on whether the witness accurately perceived the event in the first place, and whether the witness's memory degraded over time or was polluted by post-event information, jurors typically do not understand the complexity, malleability, and selectivity of memories.[70] Jurors' assessments are also subject to another cognitive error: the confidence-accuracy fallacy. Although jurors typically use eyewitness confidence as a proxy for reliability,[71] the

[64] "Elizabeth Loftus: How Can Our Memories Be Manipulated?" *NPR* (October 13, 2017), www.npr.org/transcripts/557424726 ["Manipulated"].

[65] Richard Schmechel, T. P. O'Toole, C. Easterly *et al.*, "Beyond the Ken? Testing Jurors' Understanding of Eyewitness Reliability Evidence" (2006) 46:2 *Jurimetrics* 177 ["Beyond the Ken"] at 195.

[66] Ibid.

[67] Elizabeth Loftus & Hunter Hoffman, "Misinformation and Memory: The Creation of New Memories" (1989) 118:1 *Journal of Experimental Psychology: General* 100 at 100 (noting that "postevent information can impair memory of an original event").

[68] A witness who is exposed to leading questions by investigators, recollections by other witnesses, or news reports that differ from her own memory may begin to remember the event differently in a way that aligns more closely with the narratives heard from others. According to expert Elizabeth Loftus, "[i]t's not that hard to get people to believe and remember things that didn't happen." "Manipulated", note 64 above.

[69] Elizabeth Loftus, "How Reliable is Your Memory?" (presentation delivered at TEDGlobal 2013: Think Again, June 11, 2013), www.ted.com/talks/elizabeth_loftus_how_reliable_is_your_memory.

[70] "Beyond the Ken", note 65 above, at 195.

[71] This causes jurors to "dramatically overestimate the accuracy of eyewitness identifications." Kevin Jon Heller, "The Cognitive Psychology of Circumstantial Evidence" (2006) 105:2 *Michigan Law Review* 241 ["Cognitive Psychology"] at 285; see also "Beyond the Ken", note 65 above, at 199 (31 percent of potential jurors stated a witness who was "absolutely certain"

correlation between witness confidence and accuracy is quite weak.[72] And because people tend to overestimate the reliability of their own memories,[73] witnesses are likely to be overly confident of their recollections, leading jurors to overvalue their testimony.

Robot testimony[74] would not share these vulnerabilities and may therefore be more reliable than human testimony. The common but incorrect understanding of the nature of human memory is in fact a fairly accurate representation of the way robots create memories, in that their internal decision-making systems operate much like a recording device. As a result, the information they record is verifiable and provable without additional corroboration, unlike a person's memory. Presumably, robot memory is not dynamic or suggestible. And in certain instances, a robot may actually capture a video recording of a given incident or interaction. As a result, a robo-witness's recollection of a given memory is likely to be more accurate than that of a human witness. Robot decision-making also takes into account more data than human decision-making processes can, which means a robot is capable of presenting a more thorough and accurate representation of what happened. Robot algorithms presumably would store the code from the time of the incident, recording, e.g., the environmental stimuli it perceived before making a fateful decision. In summary, robots capture more information than their human counterparts and do so more accurately, in part because they are less susceptible to post hoc manipulation or suggestibility. These advantages should enhance the accuracy of fact-finding. The potential to interrogate or challenge robot-generated evidence would depend on the nature of the robot and its memory function. For example, if a robot captures an incident by video recording, no further interpretation by third parties would be necessary. On the other hand, if the robot's "memories" take the form of algorithm sequences,

was "much more reliable" than the witness who was not, and approximately 40 percent of potential jurors agreed with the statement "an eyewitness' level of confidence in his or her identification is an excellent indicator of that eyewitness' reliability"). When evaluating the testimony of a confident witness and an unconfident witness, jurors identified the confident eyewitness as more reliable. Elizabeth Tenney, Robert J. MacCoun, Barabara A. Spellman *et al.*, "Calibration Trumps Confidence as a Basis for Witness Credibility" (2007) 18:1 *Psychological Science* 46 at 48.

[72] "Beyond the Ken", note 65 above, at 198.

[73] When asked to evaluate the reliability of their own memories, people vastly overestimated. "Beyond the Ken", note 65 above, at 196.

[74] See Chapter 8 in this volume.

then an expert would be needed to interpret that data for a lay jury, akin to interpreting DNA test results.

Furthermore, because memory formation in robots operates like a recording device, confidence may indeed be a strong indicator of accuracy in future robot testimony.[75] Because the way robots form and recall memories is more similar to the commonly held understanding of memory, people's existing heuristics are likely to help them to understand and evaluate robot testimony more accurately than human eyewitness testimony. As a result, robot witnesses ostensibly would be more reliable and improve the accuracy of litigation outcomes. A robot's internal operating algorithm may also be able to produce a confidence interval for what it saw or why it made the decision it did. Experts could then interpret and explain this confidence interval to the lay jury.

III.D.4 The Preference for Direct Evidence and Eyewitness Testimony

Despite the well-documented unreliability of eyewitness testimony, several cognitive biases cause jurors to give it greater weight than circumstantial evidence, e.g., DNA evidence or fingerprints. Because of their preference for univocal evidence requiring fewer sequential inferences, jurors typically prefer direct evidence to circumstantial evidence.[76] Combined with the misunderstanding of memory described above, these phenomena threaten the jury's fact-finding mission.

Several features that distinguish eyewitness and circumstantial evidence cause jurors to draw erroneous conclusions about their relative accuracy. First, direct testimony is told as a narrative, from a single perspective that allows jurors to imagine themselves in the witness's shoes and to determine whether the proffered explanation is plausible. As a result, jurors tend to give greater weight to direct evidence like eyewitness testimony than to highly probative circumstantial evidence, such as DNA evidence, because direct evidence requires them to make fewer sequential inferences.[77] Eyewitness testimony is, at bottom, a story: "a

[75] Cf. John Wixted & Gary Wells, "The Relationship between Eyewitness Confidence and Identification Accuracy: A New Synthesis" (2017) 18:1 *Psychological Science in the Public Interest* 10 at 55 (noting that in ideal conditions confidence level at initial identification is actually a good proxy for accuracy).

[76] "Cognitive Psychology", note 71 above, at 267–268.

[77] Ibid. at 265, 267.

moment-by-moment account that helps [jurors] imagine how the defendant actually committed it."[78] In contrast, although abstract circumstantial evidence like DNA may be statistically more reliable than eye witness testimony, it does not allow the juror to visualize an incident happening.[79] Direct evidence is also univocal; when an eyewitness recalls the crime, she speaks with one voice, frequently in a singular, coherent narrative. Circumstantial evidence, by contrast, allows for, and often requires, many inferences. In this way, it is polyvocal; multiple pieces of evidence provide different snippets of the crime.[80] Jurors must fit those pieces together into a narrative, which is more difficult than following a single witness's story. Finally, eyewitness testimony can be unconditional. An eyewitness can testify that she is absolutely certain that the defendant committed the crime, or the defendant admitted as much.[81] In contrast, circumstantial evidence is inherently probabilistic.[82]

Jurors' preference for direct evidence is driven by the simulation heuristic. The simulation heuristic postulates that people estimate how likely it is that something happened based on how easy it is for them to imagine it happening; the easier it is to imagine, the more likely it is to have happened.[83] Studies have shown that when jurors listen to witness testimony, they construct a mental image of an incident that none of them witnessed.[84] Relatedly, the ease of simulation hypothesis posits that the likelihood a juror will acquit the defendant in a criminal case depends on her ability to imagine that the defendant did not commit the crime.[85]

A variety of factors could influence how the human preference for direct eyewitness testimony would interact with robot-generated testimony. As noted above, in experimental settings participants preferred and were more readily persuaded by embodied robots that were framed in an anthropomorphic fashion, and participants preferred certain attributes like faces and a mirroring conversational style. If a robot with the preferred design gave "eyewitness" testimony, it could provide a single

[78] Ibid. at 265.
[79] Ibid.
[80] Ibid. at 267.
[81] Ibid. at 268.
[82] Ibid.
[83] Ibid. at 260.
[84] Elizabeth Loftus, "Psychological Aspects of Courtroom Testimony" (1980) 347 *Annals of the New York Academy of Sciences* 27 at 27–28.
[85] "Cognitive Psychology", note 71 above, at 262.

narrative and speak in a confident univocal voice. Assuming that the same cognitive processes that guide jurors' evaluations of direct and circumstantial evidence apply equally to such evidence, jurors would give it greater weight than circumstantial evidence. In the case of direct robot testimony, however, many of the inadequacies of human eyewitness testimony would be mitigated or eliminated altogether because robot memory is not subject to the many shortcomings of human memory. In such cases, the cognitive bias in favor of a single, confident, univocal narrative would not necessarily produce an inaccurate weighting of the evidence. However, as noted above, jurors would likely employ the same unconscious preferences for certain facial features, interaction, and speech that they apply to human witnesses.

On the other hand, robot-generated evidence not presented by a direct robo-witness might not receive the same cognitive priority, regardless of its reliability, as human eyewitness testimony. But framing and designing robots to enhance anthropomorphization, like a car with voice software and a name, might elevate evidence of this nature above other circumstantial or documentary evidence. Perhaps in this context, anthropomorphization could enhance accuracy by evening out the playing field for some circumstantial or documentary evidence that jurors might otherwise give short shrift.

III.D.5 Distributional Issues

Resource inequalities are already a serious problem in the US litigation system. Because litigation is so costly, particularly under the American Rule in which each party bears its own costs in civil litigation,[86] plaintiffs without substantial personal resources are often discouraged from bringing suit, and outcomes in cases that are litigated can be heavily impacted by the parties' resources. Parties with greater resources may be more likely to present robot-generated evidence, and more likely to have robots designed to be the most persuasive witnesses. Disparate access to the best robot technology may well mean disparate access to justice, and this problem could increase over time as robot design is manipulated to take advantage of the distortions arising from heuristics and cognitive errors. On the other hand, as robots become ubiquitous in society, access to their "testimony" may become more democratized because more

[86] John Leubsdorf, "Does the American Rule Promote Access to Justice? Was that Why It Was Adopted?" (2019) 67 *Duke Law Journal Online* 257 at 257.

people across the socioeconomic spectrum may have regular access to them in their daily lives.

IV AI Testimony in Other Legal Proceedings

In this section, we consider the impact of HRI in legal proceedings other than litigation, specifically on ADR, with a focus on arbitration, and the specialized procedures of the NTSB. We do so for two reasons. First, in the United States, litigation is relatively rare, and most cases are now resolved by some form of ADR. That is likely to be true of disputes involving robo-witnesses and evidence about the actions of robots as well. Second, these alternatives address what Sections II and III identify as the critical problem in using robot-generated evidence in litigation: the tendency of humans, especially laypersons, to anthropomorphize robots and to misunderstand how human memory functions. In contrast, the arbitration process and the NTSB's procedures assign fact-finding either to subject matter experts or to decision-makers chosen for their sophistication and their ability to understand the complex technology at issue. In this section, we describe the procedures employed by the NTSB and in arbitration and consider how these forums might address the potential distortions discussed in Sections II and III.

IV.A Alternative Dispute Resolution

One way to address the issues HRI would raise in litigation is to resolve these cases through ADR. ADR includes "any means of settling disputes outside of the courtroom," but most commonly refers to arbitration or more informal mediation.[87] Arbitration resembles a simplified litigation process, in which the parties make opening statements and present evidence to an arbiter or panel of arbiters empowered to make a final decision binding on the parties and enforceable by courts.[88] Arbitration allows the parties to mutually select decision-makers with relevant industry or technical expertise. For example, in disputes arising from an AV, the parties could select an arbitrator with experience

[87] Cornell Legal Information Institute, "Alternative Dispute Resolution," www.law.cornell.edu/wex/alternative_dispute_resolution. Mediation is an informal alternative to litigation, in which adverse parties, operating through mediators, attempt to reach a settlement.

[88] American Bar Association, "Arbitration," www.americanbar.org/groups/dispute_resolution/resources/disputeresolutionprocesses/arbitration/.

in the AV industry. We hypothesize that an expert's familiarity with the technology could reduce the effect of the cognitive errors noted above, facilitate a more efficient process, and ensure a more accurate outcome. There is evidence that lay jurors struggle to make sense of complex evidence like MRI images.[89] An expert may be able to parse highly technical robot evidence more effectively. Likewise, individuals who are familiar with robot technology may be less likely to be influenced by the anthropomorphization that may significantly distort a lay juror's fact-finding and attribution of liability.

There are reasons for concern, however, about substituting arbitration for litigation. Although arbitral proceedings are adversarial, they lack many of the procedural safeguards available in litigation, and opponents of arbitration contend that arbitrators may be biased against certain classes of litigants. They argue that "arbitrators who get repeat business from a corporation are more likely to rule against a consumer."[90] More generally, consumer advocates argue that mandatory arbitration is anti-consumer because it restricts or eliminates altogether class action suits and because the results of arbitration are often kept secret.[91]

IV.B Specialized Procedures: The NTSB

Another more specialized option would be to design agency procedures particularly suited to the resolution of issues involving robot-generated evidence. The procedures of the NTSB demonstrate how such specialized procedures could work.

The NTSB is an independent federal agency that investigates transportation incidents, ranging from the crashes of Boeing 737 MAX airplanes to run-of-the-mill highway collisions. The NTSB acts, first and foremost, as a fact-finder; its investigations are "fact-finding proceedings with no adverse parties."[92] The NTSB has the power to issue subpoenas

[89] Teneille Brown & Emily Murphy, "Through a Scanner Darkly: Functional Neuroimaging as Evidence of a Criminal Defendant's Past Mental States" (2010) 62:4 *Stanford Law Review* 1119 at 1199–1201.

[90] Stephanie Zimmermann, "Trouble with Tesla: Couple Were Sold a Damaged Car, then Told They Can't Sue," *Chicago Sun Times* (September 28, 2019), https://chicago.suntimes.com/2019/9/27/20887609/tesla-arbitration-car-damage-repair-consumer-legal-chicago-kansas.

[91] Ibid.

[92] US Code of Federal Regulations (as amended February 3, 2023), Title 49 [49 CFR], §831.4(c).

for testimony or other evidence, which are enforceable in federal court,[93] but it has no binding regulatory or law enforcement powers. It cannot conduct criminal investigations or impose civil sanctions, and its factual findings, including any determination about probable cause, cannot be entered as evidence in a court of law.[94]

The NTSB's leadership and its procedures reflect its specialized mission. The five board members all have substantial experience in the transportation industry.[95] Its investigative panels use a distinctive, cooperative "party system," in which the subjects of the investigation are invited to participate in the fact-finding process, and incidents are investigated by a panel, run by a lead investigator who designates the relevant corporations or other entities as "parties."[96] A representative from the party being investigated is often named as a member of the investigative panel to provide the investigative panel with specialized, technical expertise.[97] At the conclusion of an investigation, the panel produces a report of factual findings, including probable cause; it may also make safety recommendations.[98]

The NTSB has two primary institutional advantages over traditional litigation, institutional competency and an incentive structure that fosters cooperation. First, unlike generalist judges or lay jurors, fact-finders at the NTSB are industry experts. Second, because the NTSB is prohibited from assigning fault or liability and its factual determinations cannot be admitted as evidence into legal proceedings, parties may have a greater incentive to disclose all relevant information. This would, in turn, promote greater transparency, informing consumers and facilitating the work of Congress and other regulators.

How would NTSB respond to cases involving robot-generated evidence? Certain aspects of the NTSB as an institution may make it a more accurate fact-finding process than litigation. First, finders of fact are a panel of industry and technical experts. Using experts who have either the education or the background to fully understand the technology means that an NTSB panel may be a more accurate fact-finder. Technical competence may also be a good antidote to the lay fact-finder tendency

[93] Ibid., §831.9(a)(3).
[94] United States Code (2018), Title 49, §1154(b).
[95] Biographies for all board members can be accessed from NTSB, "Board Member Speeches," www.ntsb.gov/news/speeches/Pages/Default.aspx.
[96] 49 CFR, note 92 above, §831.8 (authority of investigator in charge), §831.11(a)(1) (designation of parties by investigator in charge).
[97] NTSB, "The Investigative Process," www.ntsb.gov/investigations/process/Pages/default.aspx.
[98] 49 CFR, note 92 above, §831.4(a)–(b).

to anthropomorphize. The NTSB panel would also benefit from having the technology's designers at its disposal, as both the designer and manufacturer of an AV could be named party participants to an investigation. Second, because the NTSB experts may have been previously exposed to the technology, they also may be less susceptible to the cognitive errors in HRI. They are more likely to understand, e.g., how the recording devices in an AV actually function, so they will have to rely less on heuristics to understand the issue and reach a sound conclusion.

On the other hand, the NTSB process has been criticized. First, critics worry that the party system may hamstring the NTSB, because party participants are often the only source of information for a given incident, although the NTSB can issue subpoenas enforceable by federal courts.[99] Second, because NTSB proceedings are cooperative, their investigations do not benefit from the vetting process inherent in adversarial proceedings like litigation. Because the NTSB cannot make rules or undertake enforcement actions, critics worry the agency cannot do enough to address evolving problems. Finally, the NTSB may not have adequate resources to carry out its duties. Although it has the responsibility to investigate incidents in all modern modes of transportation, it is a fairly small agency with an annual operating budget of approximately $110 million and about 400 employees.[100] Its limited staff and resources mean that the agency must focus on high-volume incidents, incidents involving widespread technology or transportation mechanisms.

Perhaps most important, the NTSB process is not designed to allocate liability or provide compensation to individual victims, and it is entirely unsuited to the criminal justice process in which the defendant has a constitutional right to trial by jury.

IV.C A Real-Life Example and a Thought Experiment

IV.C.1 The Fatal Uber Accident

A recent event provides a real-life example of robot-generated evidence involving the forums we have described. In March 2018, an AV designed by Uber and Volvo struck and killed a pedestrian pushing a bicycle in

[99] Jack London, "Issues of Trustworthiness and Reliability of Evidence from NTSB Investigations in Third Party Liability Proceedings" (2003) 68:1 *Journal of Air Law and Commerce* 39 at 48.

[100] NTSB, "Fiscal Year 2020 Budget Request" (Washington DC: NTSB, 2019) at 7, 28, www.ntsb.gov/about/reports/Documents/NTSB-FY20-Budget-Request.pdf.

Tempe, Arizona.[101] During that drive, a person sitting in the driver's seat, the safety driver, was supposed to be monitoring the car's speed and looking out for any hazards in the road. But at the time of the crash, the safety driver was streaming TV on their phone. The car, equipped with multi-view cameras, recorded the entire incident, including the car's interior.

The NTSB investigated the incident and concluded that both human error and an "inadequate safety culture" at Uber were the probable causes of the crash.[102] It found that the automated driving system (ADS) first detected the victim-pedestrian 5.6 seconds before the collision, initially classifying the pedestrian and her bike as a vehicle and then a bicycle, and finally as an unknown object.[103] As a result, the system failed to correctly predict her forward trajectory. The car's self-driving system and its environmental sensors had been working properly at the time of the crash, but its emergency braking system was not engaged, depending solely on human intervention.[104] Finally, Uber's automated driving technology had not been trained to identify jaywalking pedestrians; in other words, the algorithm was not programmed to register an object as a pedestrian unless it simultaneously detected a crosswalk.[105]

Local authorities in Arizona declined to criminally prosecute Uber,[106] but they did charge the safety driver with criminal negligence,[107] and at the time of writing these charges were still pending. The victim's family settled with Uber out of court;[108] there was no arbitration or mediation.

[101] Ethan Sacks, "Self-Driving Uber Car Involved in Fatal Accident in Arizona," *NBC News* (March 20, 2018), www.nbcnews.com/tech/innovation/self-driving-uber-car-involved-fatal-accident-arizona-n857941.

[102] NTSB, "Highway Accident Report: Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian" (Washington DC: NTSB, 2018), www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf at v–vi (Executive Summary).

[103] Ibid. at 39.

[104] Ibid. at v.

[105] Ibid. at 16.

[106] "Uber 'Not Criminally Liable' for Self-Driving Death," *BBC* (March 6, 2019), www.bbc.com/news/technology-47468391.

[107] Kate Conger, "Driver Charged in Uber's Fatal 2018 Autonomous Car Crash," *The New York Times* (September 15, 2020), www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html.

[108] Kiara Alfonseca, "Uber Reaches Settlement with Family of Woman Killed by Self-Driving Car," *NBC News* (March 29, 2018), www.nbcnews.com/news/us-news/uber-reaches-settlement-family-woman-killed-self-driving-car-n861131.

If the civil case against Uber had gone to trial, how would the issues we have discussed play out, and how would the resolution by litigation compare to the NTSB's investigation and findings? The vehicle's video of the incident would reduce or eliminate concerns about the accuracy of human memory. Consequently, the AV's "memory" would likely improve the accuracy of the proceeding. It is unclear whether anthropomorphization would play any role. As we understand it, the robot controlling the AV had no physical embodiment, and it was not designed to have verbal interactions with jurors or with the safety driver. There was no anthropomorphic framing such as an endearing name, assigned gender, or backstory. Thus, the jury's tendency to anthropomorphize robots would likely play no significant role in its fact-finding or attribution of liability in this specific case. In a trial, the jury's task would be to comprehend complex technical information about the programming and operation of the algorithm that controlled the car. And although jurors would have the assistance of expert witnesses, it is doubtful whether they could reach more accurate conclusions about the causes of the accident than the NTSB panel. The NTSB's panel would readily comprehend the technical information, such as why the AV mischaracterized the pedestrian and her bike as an unknown object. Moreover, the jurors, presumably more than experts familiar with the technology, might be influenced by common cognitive biases to blame the human driver more than the AV.

IV.C.2 A Thought Experiment: Litigation Involving Fully Autonomous Robotaxis

Companies like Waymo and Cruise have begun deploying fully driverless taxis in certain cities. In June 2022, Cruise, a subsidiary of General Motors and supported by Microsoft, received approval to operate and charge fares in its fully driverless, fully autonomous "robotaxis" in parts of San Francisco.[109] The conditions under which these robotaxis can operate are limited. Cruise AVs are permitted to charge for driverless rides only during night-time hours, and are limited to a maximum speed of 30 miles per hour.[110] They can, however, operate in light rain

[109] Joann Muller, "Cruise's Robotaxis Can Charge You for Rides Now," *Axios* (June 6, 2022), www.axios.com/2022/06/06/cruise-driverless-taxi-san-fransisco.

[110] As of April 2023, Cruise had applied for permission to begin testing its AVs throughout California at speeds of up to 55 miles per hour (25 mph higher). Michael Liedtke, "No Driver? No Problem. Robotaxis Eye San Francisco Expansion," *AP News* (April 5, 2023), https://apnews.com/article/driverless-cars-robotaxis-waymo-cruise-tesla-684556379bb57425c8fdf35268e8046d.

and fog, frequent occurrences in San Francisco. Waymo, an Alphabet subsidiary, began carrying passengers in its robotaxis in less crowded Phoenix in 2020, and as of April 2023 it was giving free rides in San Francisco and awaiting approval to charge fares.[111] The potential safety benefits of autonomous taxis are obvious. A computer program is never tired, drunk, or distracted. And cars like Waymo's are equipped with sophisticated technology like lidar (light detection and ranging), radar, and cameras that simultaneously surveil every angle of the car's surroundings.

How would the psychology of HRI affect fact-finding and the allocation of liability if these driverless taxis were involved in accidents? Companies designing these robotaxis have many design options that might trigger various responses, including anthropomorphic projections and responses to the performance of the algorithms controlling the cars. They could seat an embodied, co-present robo-driver in the car; its features could be designed to evoke a variety of positive responses. Alternatively, and more inexpensively, the designers could create a virtual, physically embodied driver who would appear virtually on a computer screen visible to the passengers. In either case, the robot driver could be given a name, a backstory, and an appealing voice to interact with the rider. The robotaxi driver would play the same social function as today's Uber or taxi driver, but unlike their human counterparts, the robot drivers might play no role in actually operating the vehicle.

Design choices could affect ultimate credibility and liability judgments. For example, as experimental studies indicate, giving the car more anthropomorphic qualities, a name, an appearance, a backstory, etc. would make it more likeable, and as a result, people may be more hesitant to attribute liability to it – particularly if there is a human safety driver in the car. And if both the automated car and a car with a human driver were in an accident, the experimental studies suggest that the human driver would be blamed more. The fact-finders' evaluation of algorithmic evidence might also be affected by cognitive biases, including the tendency to discount algorithmic predictions once they have been shown to be in error, even if humans have made the same error.

This example also highlights other factors that may affect the ability of various fact-finders to resolve disputes arising from the complex and rapidly evolving technology in AVs. Arbitrators vary by specialty, and some may eventually specialize in disputes involving AVs. Finally,

[111] Ibid.

the NTSB is the most knowledgeable body that could handle disputes involving AVs. However, given the structural limitations of the agency, its decisions of fault are not legally enforceable against the parties involved.

V Conclusion

Human responses to robot-generated evidence will present unique challenges to the accuracy of litigation, as well as the transparency of the legal system and the perceptions of its fairness.

Robot design and framing have the potential to distort fact-finding both intentionally and unintentionally. Robot-generated evidence may be undervalued, e.g., because it is not direct evidence. But such evidence may also be overvalued because of design choices intended to thwart or minimize a robot's liability or perceived responsibility, and thus the liability of its designers, manufacturers, and owners. Although there are human analogs involving witness selection and coaching, they are subject to natural limits, limits which largely do not apply to the ex ante design-a-witness problem we may see with robots. Additionally, cognitive biases may distort assessments of blame and liability when human and robot actors are both at fault, leading to the failure to impose liability on the designers and producers of robots.

Testing the accuracy of robot-generated evidence will also create new challenges. Traditional cross-examination is ill-suited to this evidence, which may lead to both inaccurate fact-finding and a lack of transparency in the process that could undermine public trust. Cognitive biases can also distort the evaluation of evidence concerning algorithms. The high cost of accessing the most sophisticated robots and mounting the means to challenge them can exacerbate concerns about the fairness and accuracy of the legal system, as well as accessibility to justice. Accordingly, traditional trial techniques need to be adapted and new approaches developed, such as new testimonial safeguards.[112]

But the news concerning litigation is not all bad. If it is possible to reduce the distorting effects arising from cognitive errors, robot-generated evidence could improve the accuracy of litigation, capturing more data

[112] See "Machine Testimony", note 57 above (describing the potential infirmities of machine sources, providing a taxonomy of machine evidence that explains which types implicate credibility and explores how courts have attempted to regulate them, and offering a new "vision" of testimonial safeguards for machine sources of information).

initially and preserving it without the many problems that distort and degrade human memory.[113]

Finally, alternative forums, such as arbitration and agency proceedings, can be designed to minimize the evaluation of evidence and the imposition of liability on the basis of fact-finding by individuals who lack familiarity with the technology in question.

[113] See generally Andrea Roth, "Trial by Machine" (2016) 104:5 *Georgetown Law Journal* 1245 (documenting the rise of mechanical proof and decision-making in criminal trials as a means of enhancing objectivity and accuracy, at least when the shift toward the mechanical has benefited certain interests).