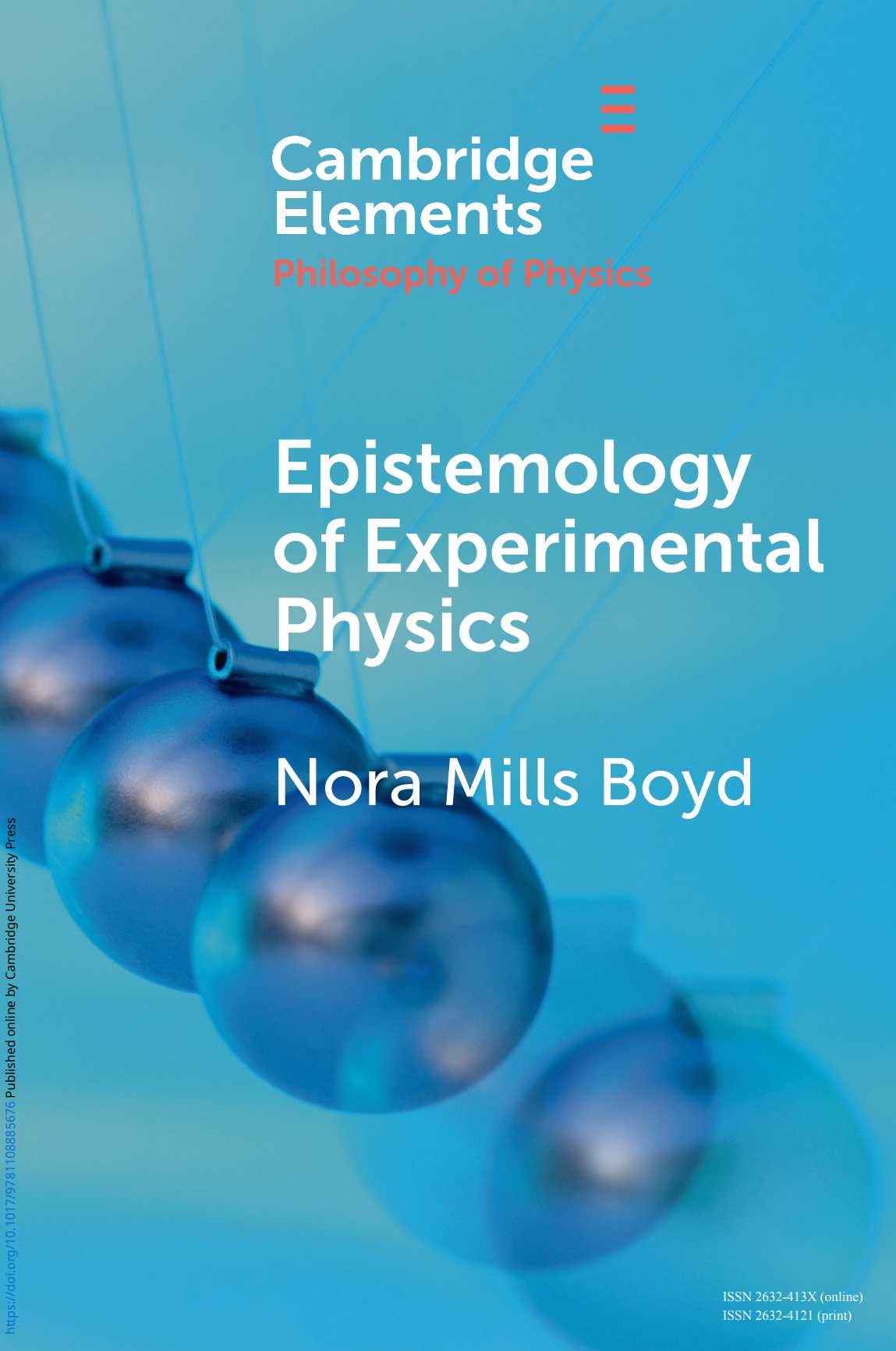

Cambridge Elements

Philosophy of Physics

Epistemology of Experimental Physics

Nora Mills Boyd

https://doi.org/10.1017/9781108885676 Published online by Cambridge University Press

ISSN 2632-413X (online)
ISSN 2632-4121 (print)

https://doi.org/10.1017/9781108885676 Published online by Cambridge University Press

Cambridge Elements

Elements in the Philosophy of Physics
edited by
James Owen Weatherall
University of California, Irvine

EPISTEMOLOGY OF EXPERIMENTAL PHYSICS

Nora Mills Boyd
Siena College

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, 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 the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108794510
DOI: 10.1017/9781108885676

First published 2021

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

ISBN 978-1-108-79451-0 Paperback
ISSN 2632-413X (online)
ISSN 2632-4121 (print)

Epistemology of Experimental Physics

Elements in the Philosophy of Physics

DOI: 10.1017/9781108885676
First published online: November 2021

Nora Mills Boyd
Siena College

Author for correspondence: Nora Mills Boyd, nboyd@siena.edu

Abstract: This Element introduces major issues in the epistemology of experimental physics through discussion of canonical physics experiments and some that have not yet received much philosophical attention. The primary challenge is to make sense of how physicists justify crucial decisions made in the course of empirical research. Judging a result as epistemically significant or as calling for further technical scrutiny of the equipment is one important context of such decisions. Judging whether the instrument has been calibrated and which data should be included in the analysis are others. To what extent is it possible to offer philosophical analysis, systematization, and prescriptions regarding such decisions? To what extent can there be explicit epistemic justification for them? The primary aim of this Element is to show how a nuanced understanding of science in practice informs an epistemology of experimental physics that avoids strong social constructivism.

Keywords: experiment, physics, calibration, experimenters' regress, commissioning

ISBNs: 9781108794510 (PB), 9781108885676 (OC)
ISSNs: 2632-413X (online), 2632-4121 (print)

Contents

1 Introduction

An epistemology in the sense intended here is the sort of philosophical beast that helps us make sense of how some form of knowledge is gained. It articulates the structures and moves appropriate for certain epistemic activities. An important lesson from philosophy of science in the twentieth century and in the first decades of the twenty-first has been that if a satisfactory epistemology of science can be provided at all, it ought to be informed by scientific practice, considered in context. For experimental physics, it has proven useful to investigate particulars regarding the experimental physicists and their skills, scientific instruments, and laboratory material culture, and how experimental work passes between various stages of research (e.g. Ackermann 1985; Chang 2004; Daston and Galison 2007; Franklin 1986, 2002, 2013; Galison 1987, 1997; Hacking 1983). In this spirit, this Element explores major issues in the epistemology of experimental physics via reflective case studies.

Experimental physics is the empirical branch of research in physics. There are many diverse subgenres, including empirical research in high-energy physics, nuclear physics, condensed matter physics, solid state physics, thermodynamics, acoustics, biophysics, astronomy, astrophysics, and cosmology. Today, physics experiments range from the homely – equipping the roofs of school buildings with cosmic ray detectors or rigging the lines trellising vintners' vines to form an ad hoc radio telescope – to the tabletop esoteric – as when physicists attempt to sequence DNA by drawing it through nanopores or use miniature explosives to detect dark matter interactions – to the opportunistic – as evidenced by the relationships between medical, energy, and defense projects and "basic research" in physics – to the behemoth multibillion-dollar infrastructure like the Large Hadron Collider (LHC) and its many associated experiments.

Indeed, "experiment" may not be the most useful category to encompass the breadth of empirical research that physicists do. For instance, physicists working in laboratories and observatories often engage in production (e.g. of the Higgs boson at the LHC) and detection (e.g. of gravitational waves using the Laser Interferometer Gravitational-Wave Observatory [LIGO]). In the present work, therefore, when I speak of physics "experiments" or "experimental physics," I intend the broader category of "empirical research in physics." Whether or not there is an epistemically distinctive role for experiment in physics, in contrast with other activities that generate empirical evidence like detection, is a topic left for another occasion.

The spectrum of epistemic successes in experimental physics is broad and impressive. The 1 part in 10^5 anisotropies in the Cosmic Microwave Background have become one of the most important sources of information in

the age of precision cosmology. The Higgs boson has been produced, gravitational waves recorded, a black hole accretion disk imaged. But the failures are diverse and fascinating as well. DAMA/LIBRA has reported a dark matter signal for decades that no other experiment has corroborated. The community at large seems suspicious of opaque and proprietary tendencies in the collaboration. BICEP2 published results claiming the discovery of the "smoking gun" signature of cosmic inflation, which by some accounts would imply the existence of the multiverse, based on rushed data analysis, and had to later retract its claim. OPERA produced results it interpreted as superluminal neutrinos, which would have catastrophic consequences for physical theory as we know it, eventually revising its interpretation after a disconnected cable was identified in its complicated apparatus.

How should we understand the reasoning exhibited in experimental physics? When does the reasoning go awry? To what extent can the epistemology of experimental physics be articulated and justified?

We would like to understand how methods employed in empirical research help us learn about the physical nature of the world. Moreover, we would like our epistemology of experiment to be in large part descriptively adequate – we want to generally capture how experimental physicists in fact act and reason so as to arrive at scientific knowledge. We want an epistemology of experiment that stays close to scientific practice, but not so close as to make it impossible to see the missteps of scientists when they happen. Our aim is not to offer a prescriptive rational reconstruction divorced from science in practice. Rather, our aim is to investigate the epistemology of experiment by intimately examining actual cases while retaining enough critical distance that we might have a hope of telling when and why things go well, and when and why they go wrong, in a manner that improves our collective understanding of the epistemic fruits science yields. In this sense, I take inspiration from the methodological approaches of Perović (2017) and Tal (2016b).

Much philosophy of science, even that which focuses on experiment in physics, has emphasized the epistemic implications of experimental results for theoretical understanding. Was the Michelson–Morley experiment decisive in overthrowing the ether? Was Eddington's eclipse expedition the impetus for the broad acceptance of relativity? Yet there is another sort of question, which is methodologically and epistemologically prior to these. When is an experimental apparatus producing good data? The success of an experiment requires sufficiently good data. But what makes data *sufficiently good*? This prior sort of question is related to that which concerned Peter Galison in his 1987 book *How Experiments End*. Galison was interested in how confidence in an experimental result, such as the discovery of a new particle, comes about in the course of an

experiment – how physicists come to be assured that "they have gold in their pans, not pyrite," "that an effect 'will not go away'" (1987, 2, 13). In microcosm, this question can be asked about a single experimental run: was everything in working order that needed to be in order to justify including the data thereby produced in the final analysis? Already this basic question houses three significant and interrelated difficulties for epistemology of experiment. Let us call them *auxiliaries*, *regress*, and *excuses*. Readers will find these three issues reverberating in the following sections. Are the data being produced in the right sort of way by the target of interest, or are any number of other possible sources of influence interfering? Is the experimental apparatus functioning properly? Should the data produced be kept and used in analysis, or are there legitimate reasons for omitting them? The ways that scientists provide answers to such questions ought to inform our epistemology of experimental physics. If "subtle contingencies," an experimenter's unspecifiable capacity for good judgment, vicious circularity, or social factors like concern for prestige, determine the outcomes of experiments, then there is little hope for an epistemology of experimental physics worth standing for.

The overall arc of this Element is as follows. Section 2 introduces the main nexus of challenges that face an epistemology of experiment via a discussion of some historically significant experiments and philosophical, historical, and sociological work that has been done on them. Two main philosophical issues surface. First is the possibility of a "crucial experiment" given the myriad auxiliary details involved in any actual experiment – that is, the problem of underdetermination in the context of experimental physics. Second is the specific epistemic challenge of instrument calibration for epistemology of experiment. The humble task of calibration has already received significant philosophical attention. This is not so surprising since in many circumstances, if the instrument to be used is not calibrated, the whole experimental show cannot go on. Yet scholarship on calibration has made it somewhat of a mystery how this early and important aspect of experimental practice is accomplished. As Harry Collins articulates it, the epistemic problem is a bad sort of circle that scientists only escape by appealing to non-epistemic resources. How can one judge that an instrument is working properly? Are there any grounds besides the sought-after successful operation on which to judge that very success? This is the problem of the "experimenters' regress" that an epistemology of experiment ought to address somehow.

Section 3 explores a third philosophical issue germane to the epistemology of experimental physics, which has received less philosophical attention than crucial experiments versus underdetermination and calibration versus experimenters' regress. This third issue concerns the appropriate omission of data. It

seems that it is sometimes appropriate to leave out certain data from the analysis. For instance, data taken during the calibration procedure, or when the experimenters had reasons to believe that the instrument was not functioning properly, may be left out of the data analysis used to produce the final result of the experiment. Such omitted data may not be considered part of the "science data" or the "experimental run." But there are clearly inappropriate reasons for omitting data too – for example, simply that the data did not agree with the experimenters' expectations, or that including them would be bad for the experimenters' public image. What counts as appropriate *excuses* for omitting data? As is perhaps already clear, the problem of determining whether data were omitted for epistemically kosher reasons can be bound up with the epistemic challenges of *auxiliaries*, and of *regress* too.

Essentially, Sections 2 and 3 explain the need to distinguish between reasons that are acceptable in the epistemology of experiment and those that do not belong. In the philosophy of experiment literature, Allan Franklin has led the way in identifying how scientists successfully argue for the significance of their results (Franklin 2016; Franklin and Perović 2019). He has written extensively about a variety of such arguments, and discussed them in the context of detailed cases. Additional work might be done in this vein by exploring whether there may be further structure to be discerned among the collection of arguments that Franklin has identified and studied. Section 4 takes a preliminary stab at organizing Franklin's epistemology of experiment. Section 4 also introduces new work in another direction, building on the calibration literature. Considered in the broad context of practices that are relevant to the epistemology of experiment, calibration can be seen as an important aspect of the larger "commissioning phase" of an experiment. Instrument calibration can occur during commissioning, perhaps many times, but calibration procedures do not exhaust all of the epistemically relevant activities of the commissioning phase. It is during commissioning that experimental physicists accomplish much of the work that they will need in order to argue for the epistemic significance of their results. Section 4 thus makes a case for further philosophical attention to commissioning as it relates to the epistemology of experiment. The concluding portion of Section 4 mentions some relevant topics that are beyond the scope of this Element for further investigation in the epistemology of experimental physics.

I have briefly mentioned the main philosophical issues that readers will encounter in the following three sections, but it may also be useful to know in advance which case studies in experimental physics will appear where. The philosophical issues of Section 2 are introduced via a discussion of Boyle's reflections on the subtle judgments involved in contextual experimental

reasoning. My purpose in including this discussion is twofold: to introduce a major figure in the history of experimental physics, and to display the recurring philosophical themes already present in this early work. Readers who have any experience working with scientific instruments and conducting experiments may well recognize familiar trials and tribulations in Boyle's reflections. Millikan's oil drop experiments and the famed 1919 eclipse expedition star in Section 3. These are both cases in which data were omitted and the rationale for so doing has been questioned. I argue against the grain of some prominent scholarship on both cases. Section 4 discusses the Large Hadron Collider and the KATRIN neutrino mass experiment, as well as some examples from metrology. Via these cases I argue that calibrating one's instrument is necessary in many experimental contexts, and need not be a circular procedure. Useful calibration procedures involve a signal other than that which is the ultimate target of the experiment (as Franklin has argued). Calibrations thus rely on epistemically prior procedures. In practice the regress thereby triggered is truncated by having adopted some tentative assumptions, which could then encounter resistance from empirical results and be subsequently refined (Chang 2004). While this does not amount to a foundationalist grounding in experience or anything else, it does show how the epistemic progress of scientific inquiry is constrained by the way the world is rather than merely by convention or speculation. Measurements, indeed instrument use in general, generate empirical data when they are properly produced by the worldly target of research. This feature is essential to the special epistemic status of science. In addition to calibration, many other checks and tests are often made in the commissioning phase of an experiment to demonstrate readiness for the purposes at hand. The outcomes of these practices need not be decided by power dynamics among people. Specific features of the experimental context can furnish good reasons for certain decisions.

2 Epistemic Challenges in Experimental Physics

> *Michelson had difficulties too. For instance, horses going by outside completely upset the experiment by the otherwise unnoticeable jiggling of the building. In the end he went to the country and floated the whole experiment in a bath of mercury to damp out the "noise."*
>
> –Ian Hacking (1983, 257)

When we survey the history of experimental physics we can discern some sticky epistemic issues that are still with us today. This section serves two functions. First, it briefly canvasses some of the historical figures and ideas typically invoked in explaining what experiments are, from whence experimental practices originated, and why they are important for advances in our theoretical

understanding of nature. Second, it exposes some of the reoccurring threads in the epistemology of experimental physics: *auxiliaries* and *regress*. We start with Francis Bacon's method of induction from natural histories, highlighting the highly influential idea of a "crucial instance," which evolved into the ubiquitous and contentious idea of a "crucial experiment." We examine some philosophical reflections on the possibility of a crucial experiment, a topic to which we return in Section 3. After Bacon, we turn to Robert Boyle's experimental methods, with special attention to the narrative presented in Steven Shapin and Simon Schaffer's iconic work *Leviathan and the Air-Pump* (2011/1985) regarding what they argue is an instance of the experimenters' regress in the matter of producing functional air-pumps.

2.1 Crucial Experiments and Underdetermination

In his seminal book *Representing and Intervening* (1983), Ian Hacking introduces Francis Bacon (1561–1626) as "the first philosopher of experimental science" (246). Hacking is not alone in this appraisal. Centering Bacon specifically in our historical narrative can be challenged (cf. Kheirandish 2009). Nevertheless, it is instructive for our purposes to take a look at what sort of methods Bacon advocated, and to what end. With the explicit aim of overhauling the approach to natural philosophy descendent from Aristotle, Bacon wrote the *Novum Organum* to promote and illustrate induction as the appropriate method for learning about nature and for putting that knowledge to practical use. For Bacon, understanding of the natural world was not to be purely "invented" or "imagined" by "premature reflection" but discovered by applying reason to thorough natural and experimental histories – collections of documented diverse instances of the matter of interest (Book II, Aphorisms IX and X). Bacon famously uses the example of the nature of heat to illustrate his approach. In investigating the nature of heat, a natural philosopher should first list many and diverse instances of heat – from lightning, to warm springs of water, to fresh horse shit – and then analyze these in comparison with nearby instances in which heat is absent (Aphorisms XI and XII). These instances could be observations of a general sort, such as that heat attends all flames, boiling liquids, and sparks struck from flint, and could also include observations made in the context of "experiments" purposefully aimed at investigating instances of heat, such as attempting to focus moonlight with a burning glass (Aphorism XII). Indeed, in his illustrative tables pertaining to heat, Bacon calls for several experiments that would need to be performed in order to fill in the details that he deems relevant to the inquiry (Aphorism XIII). After examining instances of presences and absences in these special

tables, Bacon's method recommends investigating the extent to which the nature of interest varies as a matter of degree, by making such observations as: "Animals increase in heat from movement and exercise, from wine and eating, from sex, from burning fevers and from pain" (Aphorism XIII, point 9). With all of this information set out in advance, Bacon instructs that the human mind can *then* be set to work in discerning the nature in question.

> After the presentation has been made, induction itself has to be put to work. For in addition to the presentation of each and every instance, we have to discover which nature appears constantly with a given nature or not, which grows with it or decreases with it; and which is a limitation (as we said above) of a more general nature. If the mind attempts to do this affirmatively from the beginning (as it always does left to itself), fancies will arise and conjectures and poorly defined notions and axioms needing daily correction, unless one chooses (in the manner of the Schoolmen) to defend the indefensible. And they will doubtless be better or worse according to the ability and strength of the intellect at work. And yet it belongs to God alone (the creator and artificer of forms), or perhaps to angels and intelligences, to have direct knowledge of forms by affirmation, and from the outset of their thought. It is certainly beyond man, who may proceed at first only through negatives and, after making every kind of exclusion, may arrive at affirmatives only at the end. (Aphorism XV)

Induction, then, is the necessary process by which mere mortals can hope to acquire knowledge of the true nature of things.

Bacon's process involves setting reason to the task of uncovering what the features of the elements in the collection reveal about the nature of things and the laws that govern them. Moreover, induction is an iterative process that needs rounds of refinement. Working with the tables of presence, absence, and degrees, considering how to characterize the nature of something like heat that would capture and exclude the appropriate instances, the faculty of understanding may be permitted to hazard a "first harvest" characterization. The aim is to expose the nature of heat such that it admits no contradictory instances and accounts especially for the revealing instances, *instantiae ostensivae* (Aphorism XX). Bacon's summary of his first harvest regarding the form of heat is: "Heat is an expansive motion which is checked and struggling through the particles" (Aphorism XX). From this initial attempt, the natural philosopher proceeds in induction by way of various other "aids to the intellect" (Aphorism XXI) such as considering instances of specific types in a systematic way, such as the revealing instances mentioned earlier, "which reveal the nature under investigation naked and independent" and, correspondingly, concealed instances, *instantiae clandestinae*, which "exhibit the nature under investigation at its lowest strength" (Aphorisms XXIV and XXV). The most historically notorious

of these types of instances are those Bacon calls the crucial instances, *instantiae crucis*. Bacon draws the name of this type of instance from the signposts that label different ways to go at an intersection. He describes these instances as follows:

> Sometimes in the search for a nature the intellect is poised in equilibrium and cannot decide to which of two or (occasionally) more natures it should attribute or assign the cause of the nature under investigation, because many natures habitually occur close together; in these circumstances crucial instances reveal that the fellowship of one of the natures with the nature under investigation is constant and indissoluble, while that of the other is fitful and occasional. This ends the search as the former nature is taken as the cause and the other dismissed and rejected. Thus instances of this kind give the greatest light and the greatest authority; so that a course of interpretation sometimes ends in them and is completed through them. Sometimes crucial instances simply occur, being found among instances long familiar, but for the most part they are new and deliberately and specifically devised and applied; it takes keen and constant diligence to unearth them. (Aphorism XXXVI)

In other words, when seeking an account of the nature of something and faced with two reasonable possibilities, one can attempt to find an instance that clearly displays the nature in question and reveals either of the possibilities as separable, unnecessary for that nature, such that the true cause can be ascribed to the other possibility. Bacon uses the example of the ebb and flow of the sea, which "must of necessity either be caused by" rocking as in a shallow bowl or by the rising and falling of the waters from the depths, as when boiling water rises and sinks again when the boiling subsides (Dumitru 2013, 49). Bacon suggests that a crucial instance in this context would be supplied by the observation that the high tide in Spain and Florida is simultaneous with the high tide in Peru and eastern China. Such an instance would rule out the sloshing possibility. Sometimes, Bacon says, these instances are gifted to us by nature, but more often they follow from investigation designed for that purpose.

A simplified distortion of this reasoning, inspired by but not faithful to Bacon's original, has come to be known as a "crucial experiment" (see Hacking 1983, 249–251). The caricature instructs us to sufficiently narrow the research question such that carefully constructed circumstances should yield a decisive answer. The problem, of course, with that is this: How can one be assured that the possibilities considered are exhaustive? If the possibilities considered are not exhaustive, then the purported crucial instance or the result of the purported crucial experiment will not be decisive after all. The case of the ebb and flow of the tide, for one, does not seem watertight. Are there any true crucial instances or true crucial experiments at all?

As Pierre Duhem lucidly explained, when an experiment fails to produce a predicted phenomenon, "The only thing the experiment teaches us is that among the propositions used to predict the phenomenon and to establish whether it would be produced, there is at least one error; but where this error lies is just what it does not tell us" (1991/1954, 185) and "the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed" (187). Indeed, Duhem goes so far as to embrace holism regarding physical theory:

> People generally think that each one of the hypotheses employed in physics can be taken in isolation, checked by experiment, and then, when many varied tests have established its validity, given a definitive place in the system of physics. In reality, this is not the case. Physics is not a machine which lets itself be taken apart; we cannot try each piece in isolation and, in order to adjust it, wait until its solidity has been carefully checked. Physical science is a system that must be taken as a whole; it is an organism in which one part cannot be made to function except when the parts that are most remote from it are called into play, some more so than others, but all to some degree. If something goes wrong, if some discomfort is felt in the functioning of the organism, the physicist will have to ferret out through its effect on the entire system which organ needs to be remedied or modified without the possibility of isolating this organ and examining it apart. The watchmaker to whom you give a watch that has stopped separates all the wheel-works and examines them one by one until he finds the part that is defective or broken. The doctor to whom a patient appears cannot dissect him in order to establish his diagnosis; he has to guess the seat and cause of the ailment solely by inspecting disorders affecting the whole body. Now, the physicist concerned with remedying a limping theory resembles the doctor and not the watchmaker. (187–188)

For Duhem, there can be no "crucial experiment" in physics because having one would require the physicist to "enumerate completely the various hypotheses which may cover a determinate group of phenomena; but the physicist is never sure he has exhausted all the imaginable assumptions" (190). Logic alone does not preclude different physicists from responding to mismatch between theoretical predictions and the results of experiments in diverse ways – one could choose to cut out the heart of the theory, while another could choose to pursue more superficial modifications. However, Duhem stresses that the loose directive logic furnishes in such circumstances is not the only stricture the physicist obeys. In addition there are "motives which do not proceed from

logic and yet direct our choices, these 'reasons which reason does not know' and which speak to the ample 'mind of finesse' but not to the 'geometric mind,' constitute what is appropriately called good sense" (217). Of this "good sense" Duhem writes:

> But these reasons of good sense do not impose themselves with the same implacable rigor that the prescriptions of logic do. There is something vague and uncertain about them; they do not reveal themselves at the same time with the same degree of clarity to all minds. Hence the possibility of lengthy quarrels . . . In any event this state of indecision does not last forever. The day arrives when good sense comes out so clearly in favor of one of the two sides that the other side gives up the struggle even though pure logic would not forbid its continuation . . . physicists may hasten this judgment and increase the rapidity of scientific progress by trying consciously to make good sense within themselves more lucid and more vigilant. Now nothing contributes more to entangle good sense and to disturb its insight than passions and interests. Therefore, nothing will delay the decision which should determine a fortunate reform in a physical theory more than the vanity which makes a physicist too indulgent towards his own system and too severe towards the system of another. We are thus led to the conclusion so clearly expressed by Claude Bernard: The sound experimental criticism of a hypothesis is subordinated to certain moral considerations; in order to estimate correctly the agreement of a physical theory with the facts, it is not enough to be a good mathematician and skillful experimenter; one must also be an impartial and faithful judge. (217–218)

Yet, for those interested in the epistemology of experimental physics, such an appeal to the significance of "vague and uncertain" inclinations and the need for moral fortitude may be unsatisfying. It is well and good, perhaps, to entreat experimental physicists to render in themselves a lucid and vigilant sense of how to proceed in the face of underdetermination, to cultivate their characters so as to become impartial and faithful judges. But how exactly? Once cultivated, what does such good sense recommend? The epistemology of experiment would be more transparent if we could fill in the details of Duhem's "good sense" somewhat more. Transparency is especially desirable here because without it, in the murky realm of unarticulated sensing, it is more difficult to ward off the suspicion that non-epistemic interests are playing the decisive role.

Claudia Dumitru offers Robert Boyle's (1627–1691) inaugural use of the phrase "experimentum crucis" in his 1662 *A Defence of the Doctrine Touching the Spring and the Weight of the Air* as "an almost perfect example of a Baconian crucial instance" (2013, 55). The example in question is worth knowing, and will be of further use to us in what follows. Take a tall glass vial and fill it with liquid mercury. Upend the tube and, keeping the open end momentarily well

sealed, place it oriented vertically in a dish of more mercury. Remove the seal and observe the level of the mercury in the upside-down tube fall somewhat, although not entirely out of the tube. This is a Torricellian tube. In Boyle's day, this phenomenon, the space made at the top of the inverted tube under such circumstances, attracted competing explanations. Boyle's preferred hypothesis was that it was the "spring of the air" surrounding the tube keeping the mercury up. In sketch, the competing explanation was that the would-be vacuum in the Torricellian space was so unnatural that some other subtle fluid must be present in that space, keeping the mercury up. The experiment that Boyle refers to as an "experimentum crucis" is the one Blaise Pascal arranged to be carried out by his brother-in-law, who brought a tube prepared in this manner up the Puy-de-Dôme (a lava dome nearly 1,500 meters high in central France) in stages, the mercury falling as he went. Remarking on this experiment, Boyle wrote: "since this noble phenomenon seems to follow from ours and not upon our author's hypothesis, it seems to determine the controversy" (quoted in Dumitru 2013, 55). In the next section we will see that the matter is more subtle.

2.2 Calibration and Regress

The final paragraph of Shapin and Schaffer's well-known book *Leviathan and the Air-Pump* foreshadows much of the debate between philosophers interested in the epistemology of science and sociologists of science at pains to dismantle any preferential treatment for science qua knowledge-producing activity over and above other human pursuits. Shapin and Schaffer examine the conflicting perspectives of Boyle and Thomas Hobbes with respect to the nature of philosophy and the role of experimentation. To simplify, they portray Hobbes as an antagonist of experimental methods, highlighting his efforts to undermine the trustworthiness of Boyle's experiments with the air-pump by calling attention to its propensity to leak. Shapin and Schaffer ultimately conclude that revealing the labor involved in arriving at experimental results undermines their authority: "As we come to recognize the conventional and artifactual status of our forms of knowing, we put ourselves in a position to realize that it is ourselves and not reality that is responsible for what we know. Knowledge, as much as the state, is the product of human actions. Hobbes was right" (2011/1985, 344).

These final words of Shapin and Schaffer's book neatly express a position diametrically opposed to my overall message in the present work. Shapin and Schaffer's view is that seeing the scientific sausage being made, so to speak, undermines the credibility of scientific claims. I think the opposite is true: having a detailed and nuanced understanding of scientific practice helps us appreciate the justification for science's special role as our best route to learning

about nature. There is an important caveat. Someone under the mistaken impression that science proceeds according to the algorithm presented in introductory textbooks could understandably be thrown off course by learning about actual scientific methods. But the correct lesson to draw in that case is not that scientific results are *merely* the products of human actions, and thus not to be taken especially seriously, but that the surprised person had an overly idealized view of what science is all about.

It seems to me that the biggest threat to the epistemology of experiment is encapsulated in what Harry Collins has called the "experimenters' regress." Any empiricist epistemology of science (as opposed to a rationalist one) bestows a key role on empirical results. In order to argue that the results of one's experiment are to be taken seriously, one has to argue that the experimental apparatus is working properly, but to argue that the apparatus is working properly – so says the regress – one demonstrates that the apparatus has produced the expected "correct" results (Collins 1992/1985, 84). Collins argues that it is in virtue of this looping in experimental reasoning that social factors like power, prestige, and personality get traction in decisions about how the results of experiments are interpreted and received by researchers and the scientific community at large. If Collins is right about the existence of this sort of regress, then the epistemology of experiment faces a serious problem. The experimenters' regress cannot be abolished by good research ethics alone. All fraud could cease, but social factors would still be required to truncate the regress. If social factors were to play the decisive role in the deliverances of science, then there is not much to recommend a special epistemic status for the enterprise. Shapin and Schaffer explicitly invoke Collins's notion of the experimenters' regress in their discussion of attempts to replicate Boyle's experiments with the air-pump (2011/1985, 226). In order to replicate Boyle's experiments, other experimenters had to be able to judge when such replication had been accomplished. The only way to do this was to use Boyle's phenomena as *calibrations* of their own machines. To be able to produce such phenomena would mean that a new machine could be counted as a good one. Thus, before any experimenter could judge whether his machine was working well, he would have to accept Boyle's phenomena as matters of fact. And before he could accept those phenomena as matters of fact, he would have to know that his machine would work well (226).

In their narrative and analysis of these replication attempts, Shapin and Schaffer emphasize the role of social conventions for decision-making in these experimental contexts. In particular, they argue that social conventions played a role in when an apparatus under construction was accepted as an air-pump, when the results of trials made with that apparatus were taken seriously,

and when the results of trials made by anyone besides Boyle himself would be believed over Boyle's reports. If Shapin and Schaffer are correct that Boyle's experiments with the air-pump suffer from the experimenters' regress, then the epistemology of Boyle's experimental methods would be dubious. Are they correct? To answer this question, it will be helpful to examine Boyle's approach in a little more detail.

Robert Boyle is widely regarded as one of the most significant founders of experimental methods, if not *the* primary founder. For instance, writing of Boyle's experiments with the air-pump, Michael Nauenberg states: "The construction of the vacuum pump and the discovery of the physical properties of air were, undoubtedly, the most important development in experimental science at the beginning of the scientific revolution" (2015, 330–331). Taking up the project outlined by Bacon, Boyle set about collecting observations and performing experiments in order to amass empirical reserves that could serve to inform theoretical understanding of nature and useful practical applications. Boyle was evidently humble enough to suppose that great theoretical systems take significant time to emerge from natural and experimental histories, a process that might outlast his own lifetime. He saw himself as mining the marble from the quarry from which others might sculpt masterpieces later on (Sargent 1994, 64; see also Sargent 1995). Peter R. Anstey's (2014) account, which groups the approaches of Bacon-Boyle-Hooke (BBH) into a single philosophy of experiment, emphasizes that the experimental philosophers were keen to maintain healthy intellectual distance between their observations and experiments on one hand, and speculation on the other, while nevertheless recognizing the eventual aim of systematizing the results of accumulated investigations:

> On the BBH view, the overarching framework for understanding the relation between experiment and theory is that of the construction of natural histories while at the same time keeping the chief speculative theories in view. The experimenter collects observations and experiments pertaining to one of the major titles of natural history, say a particular quality such as cold or colour, and this process is loosely guided by an awareness of the leading speculative systems. But in order to avoid the dangers of prepossession the experimenter is not to engage too closely with the speculative theory. This is the background to Boyle's remark . . . "my backwardness to frame Theories has made me chuse to forbear as yet to methodize them." Likewise, Hooke claims that the experimenter "ought to be free from dogmaticallness & prejudice and not wedded to this or averse to that Opinion." It is not until the natural historical stage is well advanced, a task too large for any single experimenter, that the natural philosopher is in a position to 'methodize' their observations and experiments. (116)

Indeed, Anstey notes that a "significant portion of Boyle's experimental work had no immediate bearing on speculative theory whatsoever" (122). Something of Boyle's methods and personality can be gleaned from reading his detailed and entertaining reports of his investigations. Take for instance "Some Observations about Shining Flesh," a letter that Boyle presented to the Royal Society. Boyle recounts how he had been getting ready to go to bed when he was informed that one of his servants had found a shining neck of veal in the larder. Despite having a cold (from experimenting with a new telescope earlier that day in high winds), he stayed up to make several observations and experiments that very night, although not "long enough to make all the tryals that I thought of and judg'd the occafion worthy of" (1672, 5108). Boyle explains the origin of the veal, the number of shining spots on it, the size and shape of the spots, where they were located on the piece of meat, the strength of the light as compared to that of glow worms, the color of the light, the lack of accompanying heat, the state of the meat (not stinky), the condition of the larder in which the veal neck had been kept, including the orientation of the window thereof, the prevailing meteorological conditions such as the direction and character of the wind, temperature, quarter of the moon, and the reading of the barometer. He investigates the effect of removing a shiny part from the larger neck piece. He tries rubbing it on his hand, compressing it, placing it in a crystal vial of Spirit of Wine, in a teacup of cold water.

Then he calls for the Pneumatical Engine (the air-pump) to be set up, puts one of the largest shining pieces in the device's receiver, and has the pump worked in the dark so as to observe how the shining meat behaves as the air is evacuated from the chamber.

When the light does not appear to diminish to any considerable degree, Boyle wonders if the air-pump, "having been managed in the dark, had leaked all the while" (5112). The candles are brought back in, a Mercurial Gage is procured, the seal between the glass of the receiver and the engine is reestablished securely, and the trial is performed again. This time, Boyle witnesses the diminution of the shining light with the evacuation of air from the receiver (which evacuation is corroborated by the appearance of the gauge). When air is let back in the receiver, the increased luminosity of the piece of veal returns to its former strength. The experiment is repeated and Boyle wonders if the shining light would vanish entirely if the pumping continued long enough, but eventually he feels it is unreasonable to stay up any longer that night and that he really should go to bed.

However, while he is undressing, it occurs to him to see if other parts of the veal are "innobled" with the strange shine and, sure enough, a leg is brought to his chamber! Boyle then explains that the next day he was distracted from the

shining veal by the illness of his niece, and by the time he was disposed to perform further observations and experiments, the veal of interest had been disposed of, leaving him only the little piece preserved in the crystal vial, which he observed for several more days until the glowing diminished to imperceptibility, along with a piece of fowl with some shining parts (although not as many as the veal) that was also retrieved from the larder later.

Clearly, Boyle's methodological approach reached far beyond recording careful observations, although he certainly did that too. Taking cues from craftsmen and tradesmen, he contrived many artificial circumstances, manipulating natural conditions in order to perform particular experiments of interest. As Sargent explains, Boyle clearly recognized the difficulties that come with preparing experiments in this way:

> Although the artifice employed by practitioners provided Boyle with numerous examples of the virtues of experimental manipulation, he also learned that experiments are subject to a number of problems that only a deep and prolonged exposure to the practice would reveal. According to Boyle, experiments are "seldom solitary." The complexity of the experimental method, which makes artificial manipulation especially suited for discovering the complexity of natural processes, also gives rise to a number of problems. (1994, 69)

Subtle differences in experimental conditions can affect the success of the trial as whole, as Boyle observed from attending to the practices of smiths and glassworkers, noting, for instance, that "none but an artist expert in tempering of iron would suspect that so small a difference of time of its stay in the flame could produce so great a difference in its tempers" (quoted in Sargent 1994, 70). Boyle's awareness of the subtleties and fickleness of experiments can be easily discerned from his two fascinating essays "Of the Unsuccessfulness of Experiments" and "Of Un-succeeding Experiments." In the first, Boyle recounts that he has learned that experiments that succeed once, but that the diligent experimenter cannot manage to replicate, may fail on account of the materials employed. He complains at length of the difficulties involved in procuring Spirit of Salt unadulterated by Spirit of *Nitre*, even from supposedly reputable Chymists. Indeed, Boyle goes on to explain that even when one has managed get one's hands on what really ought to be some pure material, one may still encounter variation that can affect the outcome of experiments. Urine, for instance, which was a commonly used ingredient, differs according to its origin such as between "that of healthy and young men abounding much more with volatile Salt than that of sickly or aged persons; and that of such as drink Wine freely being much fuller of spiritous and active parts than that of those whose drink is but Beer or Water" and how long it has been kept before use in

experiments (Boyle 1999/1661, 49). He reports "great odds there may be betwixt some Experiments made with recent and putrifi'd Urine," as evidenced by the properties of a stock of urine that he had buried in a dunghill and, having been distracted by other things, left for four or five months instead of the intended five or six weeks (50). Assuming consistency among mineral samples can be particularly troublesome. He remarks that "there is as well a difference in Minerals of the same kind, as there is in Vegetables and Animals of the same species," owing to the conditions of the minerals' origins, which means that what appears to be a pure sample "may have lurking in them Minerals of quite other nature, which may manifest themselves in some particular Experiments" (46). Of course, prepared materials can exhibit diverse effects on experiments as well, depending on the shape of the glassware, for instance, and can even be over-purified for the purpose at hand:

> For instance, we have sometimes for recreation sake, and to affright and amaze Ladies, made pieces of white paper and linnen appear all on a flame, without either burning, sindging, or as much as discolouring them. This is performed by plunging the paper very thoroughly in weak Spirit of Wine, and then approaching it to the flame of a candle, by which the spiritous parts of the Liquor will be fired, and burn a pretty while without harming the paper. But if this Experiment be tryed with exquisitely rectifi'd Spirit of Wine, it will not succeed. (53)

In the latter essay on un-succeeding experiments, Boyle discusses other contingencies, circumstances, and abstruse causes, besides the variety of materials that can spoil experiments. The purpose of the essay is twofold. First, Boyle aims to impress upon other experimenters examples that should inspire them to "try those Experiments very carefully, and more than once, upon which you mean to build considerable Superstructures either theoretical or practical, and to think it unsafe to rely too much upon single Experiments, especially when you have to deal in *Minerals*" (77). Sometimes an experiment that works on a small quantity of matter will fail on a greater quantity. An experiment performed at one time (such as certain dissections of animals) may not produce the same results performed in another time of the season. Whether, as Bacon apparently himself reported, roses can be made to bloom again in autumn may depend both on the kind of rose and on the prolific nature of the individual bush (59). Moreover, since even the most skilled craftsmen, like dyers and glassworkers, can sometimes fail to reproduce some technique or material that is exceedingly familiar to them, "it need be no such wonder, if Philosophers and Chymists do sometimes miss of the expected Event of an Experiment but once, or at least but seldom try'd, since we see Tradesmen themselves cannot do *always*, what, if they were not able to do *ordinarily*, they could not earn their bread" (64–65). In

general, the particular circumstances of an experiment that "are very difficult to be observed, or seem to be of no concernment to an Experiment, may yet have a great influence on the Event of it" (65). There are issues with the testimony of experimenters too. Boyle suggests that the reported results of some experiments depend very much on the pride of the experimenters who perform them, such as reports of certain plants frozen in ice: "'tis strange to observe what things some men will fancy, rather than be thought to discern less than other men pretend to see" (62). Boyle cautions experimenters not to be lazy, of course, in attempting to discern the causes of varied experimental outcomes (as when differences in the buds of grafted cherry trees foretell whether they will fruit the first or second year after grafting). Nevertheless, the essay is full of examples where subtle differences in materials and circumstances evade and frustrate even the careful experimenter for some time.

Second, the essay is meant to serve as a kind of "Apology for Sober and Experimental Writers" in the event that one attempts to perform an experiment described by another experimenter and the trial fails to produce the anticipated result (77). When these failures occur, the reputation of the writer should not be immediately and utterly undermined, since perhaps subtle differences in the circumstances of the experiment are at fault. Along the same vein, Boyle intimates that when reading the reports of an experimenter he does trust, even if the reported results contradict other observations or widely held hypotheses, he will not immediately discount the results since "*sometimes* there happen irregularities contrary to the usual course of things" and sometimes the contradiction is only apparent (80). In short, Boyle cautions his fellow experimenters to be watchful and wary, but not to despair so much as to forsake the endeavor because even in the mistakes there can be interesting discoveries.

According to Sargent, Boyle regarded the results of experiments as "merely signs that provide hints about how nature works" (1994, 71), and he accepted theories that enjoyed the support of many complementary experiments:

> Boyle developed his epistemological criterion for the acceptance of theoretical claims in a manner consistent with the way in which he constructed his experimental philosophy in general. A concurrence of probabilities is achieved when "the most information procurable that is pertinent to the things under consideration" supports a particular conclusion and there is no evidence to the contrary that would militate against it. Concurrence is a somewhat vague epistemological criterion, however, in large part because it is based upon considerations of relevancy that in turn will depend upon the current state of knowledge concerning a particular subject matter. Boyle was aware of the fact that such a "judgment of reason" would therefore be context-dependent, and that the acceptability of knowledge claims could change as the context of reason changed. Experimental proof is a complex

> and dynamic process. For this reason, he wrote that he would "not debar" himself from revising his opinions in the event that "further progress in history shall suggest better hypotheses or explications" (65)

In particular, due to the multifarious subtle factors involved in even the most careful experiments, a single experiment should not be taken as proof nor as refutation of a theoretical claim (71).

Other natural philosophers wanted to build their own air-pumps and to perform experiments and explore phenomena to which Boyle had drawn attention. Partly because the construction of an air-pump was an expensive endeavor, there were few initial successes in producing rival apparatuses. The pumps themselves were temperamental, and as Shapin and Schaffer stress, prone to leaking, which we have already seen from Boyle's own observations on the shining veal. Much care and tinkering was therefore required in order to construct an operational pump suitable for investigation of phenomena. Shapin and Schaffer describe in detail the attempts of a particular challenger of Boyle's who created his own pump based on details of Boyle's pump as conveyed by Robert Moray. According to Shapin and Schaffer, "Christiaan Huygens was the only natural philosopher in the 1660s who built an air-pump outside of the direct management of Boyle and Hooke" (2011/1985, 235). In the course of attempting to build his own serviceable pump, Huygens introduced several modifications to Boyle's design. For instance, Huygens's pump had a copper valve where Boyle's had a wood one, had a closed top whereas Boyle's had a port, and used different sealant materials (237). Like Boyle, Huygens made subsequent incremental alterations to his apparatus in an attempt to improve its functioning. Huygens told others that his pump was better than Boyle's and offered as evidence an inflated bladder that remained inflated in the pump all night, whereas Boyle had reported that bladders deflated (slowly) in his apparatus (237). Once he had his pump running to his own satisfaction, Huygens claimed to have made an interesting discovery regarding the Torricelli-type experiments that Boyle described in *New Experiments*. Placing a barometer containing "water he had purged of air by leaving it many hours in the receiver of the air-pump," he discovered that, unlike ordinary water, the level of this water did not fall with the evacuation of the pump (241). According to Shapin and Schaffer, Huygens came to use this "anomalous suspension" as a calibration phenomenon (243). "By February 1662 Huygens took the anomalous suspension of water well purged of air, and the fall of that water when a bubble was introduced, to be marks of a good machine" (243).

The anomalous suspension was a result unforeseen by Boyle, and not immediately replicated by him either. When news came from Holland of this strange discovery, the pumps in England were not in a position to verify the finding –

one was under development and the other was temporarily out of commission (244). Without appropriate trials of his own, Boyle had to respond to Huygens. According to Shapin and Schaffer, "Boyle claimed that his machine was better than that of Huygens, so that anomalous suspension was a mark of Huygens' *incapacity* to make a good pump," "denied that anomalous suspension *could* be used as a calibration of the pump," and even suggested to Huygens (via Moray) that the latter's apparatus was probably not sufficiently evacuated and that Huygens should employ a gauge to check (245). Huygens persisted in defending the phenomenon and providing further results to support it, and Boyle continued to blame the anomalous suspension on the integrity of his challenger's pump (246).

Despite further technical improvements, trials, and arguments from Huygens, by January 1663 Boyle was still not "assured of the truth of the experience" (248). According to Shapin and Schaffer, "in March and April 1663 it became clear that unless the phenomenon could be produced in England with one of the two pumps available, then no one in England would accept the claims Huygens had made, or his competence in working the pump" (249). Failing to sufficiently purge water of air for use in the experiment, Moray even ventured that the difference in experimental outcomes might be due to the water in London being different than the water in Holland (249). Shapin and Schaffer argue that unless the anomalous suspension could be reproduced on Boyle's turf and to his satisfaction, they were at a stalemate. However, continued technical difficulties in England prevented them from performing satisfying trials. Shapin and Schaffer describe the status of the pump that Boyle had left to the Royal Society, which was put under Robert Hooke's charge in late 1662, as "an almost permanent trouble because of its obvious leakage" (249). Eventually, in mid-1663, Huygens made the trip to London. Boyle was out of town at the time, residing at his sister's house in Essex. Within about a week of Huygens's arrival, Hooke had successfully produced the anomalous suspension (251). Boyle was informed straightaway and witnessed the phenomenon himself in experiments performed after he returned to London in August (252). Nevertheless, Boyle maintained that "in regard they had noe Gage to try how farre they had exhausted ye Aire in the Receiver it seem'd not absurd to coniecture that there might remaine in ye Receiver enough [air] to keep up in ye Tube 3 or 4 foot of Water" (quoted in 252). In other words, even after seeing the anomalous suspension himself, Boyle was not totally satisfied that it was *not* due to a problem with the instrument. According to Shapin and Schaffer, in experiments made during the fall of 1663, Boyle attempted to dissociate the troublesome phenomenon from the appraisal of a pump's integrity by producing the anomalous suspension of mercury outside of the air pump entirely (254).

Shapin and Schaffer argue that Boyle never published an account of anomalous suspension because this phenomenon "resisted the recognized explanatory competence of the spring of the air" and "might challenge . . . the worth of the air-pump" (230, 255). Shapin and Schaffer take this episode to support the main interpretive thrust of their book – the intimate connection between the "problem of knowledge" and the "problem of social order":

> The career of the air-pumps in the 1660s shows how experimenters made matters of fact. Two points can be made: (1) the accomplishment of replication was dependent on contingent acts of judgment. One cannot write down a formula saying when replication was or was not achieved. The construction of any device which could be taken as a successful copy of an existing pump was entirely dependent on direct witnessing . . . (2) Thus, if replication is the technology which turns belief into knowledge, then knowledge-production depends not just on the abstract exchange of paper and ideas but on the practical social regulation of men and machines. The establishment of a set of accepted matters of fact about pneumatics required the establishment and definition of a community of experimenters who worked with shared social conventions: that is to say, the effective solution to the problem of knowledge was predicated upon a solution to the problem of social order. Hobbes's criticism was that no matter of fact made by experiment was indefeasible, since it was always possible to display the labour expended on making it and so give a rival account of the matter of fact itself. The decision to display or to mask that labour was a decision to destroy or to protect a form of life. (281–282)

Are Shapin and Schaffer right to interpret Boyle's judgments on the anomalous suspension as primarily a matter of protecting his preferred hypothesis and his pride in the worthiness of his instrument? Does the anomalous suspension display the decisive role of social convention in settling when air-pumps were working properly and thus when the results of trials made with them ought to be taken seriously?

Interestingly, the issue of anomalous suspension remained mysterious long after the height of the controversy between Boyle and Huygens. In a 2015 article, Nauenberg states that the origin of the phenomenon "was not clarified, and it has remained an unsolved puzzle to the present day" (329). Nauenberg actually undertook to perform and compare the Boyle-Hooke and Huygens versions of the experiment to try to understand the physics from a contemporary perspective. He found that the explanation Shapin and Schaffer offered was largely wrong (340). Using ordinary tap water, Nauenberg saw the water column in his Torricellian tube fall upon the commencement of pumping, which he attributes to "the pressure of the gas formed by the bubbles surfacing inside the tube" (337). When the experiment

was performed with water purged of some of its dissolved air, the column of water did not start to fall until the pressure in the receiver was much lower than it had been with the ordinary, unpurged water (337). According to Nauenberg, the anomalous suspension is due to the limitations of the seventeenth-century pumps to evacuate their receivers – he estimates that there would have been a residual air pressure greater than 0.15 atmospheric pressure (336–337). Indeed, he argues that Boyle himself was aware of the air bubble issue (335–336). That is, purged water could have remained anomalously suspended in these early pumps because the pumping mechanism simply was not strong enough to sufficiently remove residual air pressure in the receivers to draw down a column of water without the help of the pressure of a gas bubble released inside the tube itself. Since Nauenberg used a modern vacuum pump, it would be interesting to try these experiments with more faithful replicas of the seventeenth-century instruments, wax and resin seals and all.

As we have seen from Boyle's two essays on "unsuccessful" and "unsucceeding" experiments, which were originally published prior to the Huygens controversy, Boyle was very aware of the difficulties involved in reproducing an experiment and the many subtle circumstantial differences that could disrupt any particular trial. Boyle need not have been dogmatic to be suspicious of Huygens's results. Huygens was using an instrument of his own construction with a newly modified design. Boyle knew very well how temperamental his own air-pumps could be. In fact, the reason he could not immediately investigate the anomalous phenomenon was that his pumps were out of commission. Even in low-stakes trials, such as the late-night experiments with the shining veal neck, the seal on the air-pump required constant vigilance. It was not unreasonable for Boyle to suspect that Huygens's early experiments with designing, building, and operating an air-pump would meet with difficulties, and that Huygens's initial reports of unusual results should be interpreted cautiously. It was reasonable for Boyle to worry that Huygens's pump leaked. Even after seeing the experiment performed himself, it was not necessarily dogmatic for Boyle to worry about air being left in the apparatus. This was not special treatment for Huygens – this was Boyle's cautious attitude about the functioning of these temperamental instruments born of hard-won experience, and a wariness consistent with the general attitude toward experiments expressed in the two essays on the unsuccessfulness of experiments. A multitude of contingencies can affect the outcome of any one experiment.

From the perspective of the experimenters' regress, however, it is perhaps little comfort to appeal to the reasonableness of caution in light of subtle contingencies. In a given experiment, one wants to know whether those

contingencies have spoiled the results for the purpose at hand. One wants to know if the apparatus is working properly. If one can appeal to "subtle contingencies" ad nauseum, then it seems that convention or other social influences must step in to decide the verdict. Thus we see again the connections between underdetermination and the experimenters' regress in the epistemology of experiment – the connections between the question of auxiliaries and whether an apparatus is working as the experimenter desires. In the following section, we extend this discussion to our third, related issue: *excuses*. Ordinarily, it is not epistemically permissible to ignore empirical data because they defy one's expectations. But data are routinely omitted on the grounds that the experimental apparatus was not working properly at the time the data were generated. If these judgments are decided by social factors too, then much of experiment would be just as good as fraud.

3 Epistemology of Data Omission

> *Let us admire them as craftspersons: the foremost experts in the ways of the natural world.*
>
> –Harry M. Collins and Trevor Pinch (1993, 142)

There are certain physics experiments that are often presented in pedagogical contexts, but whose pedagogical virtues for physics students derive from distortions of the historical methodological details: Galileo's inclined plane, Poisson's white spot, the Michelson–Morley experiment, the Millikan oil drop experiments. Indeed, part of the education of a physics student today often involves performing versions of these very experiments in laboratory courses – not reenacting them as the historical figures would have performed them, but rather executing cleaned-up, sometimes even prepackaged remixes of the classics. This approach risks transmitting an unrealistic expectation about the capacity for individual experiments to be revolutionary, surgical, and decisive – and, even more unrealistic still, to come with instructions.

Due in part to this disconnect between the historical and pedagogical features of certain experiments, such experiments are also canonical in the academic literature on the epistemology of experiment. Excellent history and philosophy of science scholarship on several canonical experiments often reveals a common characteristic – on the surface, these experiments lend themselves to neat philosophical storytelling; however, historical and contextual investigation complicate and enrich the role of these cases in our evolving understanding of physics (see e.g. Brush 1999, Holton 1969, and Worrall 1989).

Some of the fascinating nuances of experiment in practice surround the issue of selectively choosing which of the data collected in the course of an experiment

will actually be used in the final analysis (see Franklin 2002). Leaving data out of an analysis without just cause falls among what have come to be known as "questionable research practices" (QRPs), particularly in the context of the replication crisis in psychology and beyond, *p* hacking being among the most widely recognized of such practices. Indeed, Hannah Fraser and colleagues (2018) characterize *p* hacking as a family of related practices that involve making data collection and/or inclusion decisions based on the significance of the result thereby obtained: "checking the statistical significance of results before deciding whether to collect more data; stopping data collection early because results reached statistical significance; deciding whether to exclude data points (e.g. outliers) only after checking the impact on statistical significance and not reporting the impact of the data exclusion; adjusting statistical models, for instance by including or excluding covariates based on the resulting strength of the main effect of interest; and rounding of a *p* value to meet a statistical significance threshold (e.g., presenting 0.053 as P < .05)" (2). By *p* hacking, researchers let their desire for a publishable statistically significant result get the better of their fidelity to what information the data have to offer about the subject matter of interest. Such practices are epistemically detrimental because they rob us of opportunities to learn from empirical research. It would clearly be antithetical to the aims of science for researchers to simply choose which data to include in analysis at will. Such a practice would allow scientists to ignore "inconvenient" anomalies and thus make agreement between theory and evidence uninformative. Anyone with minimal empiricist commitments should worry about the prevalence of such practices in science (Boyd 2018a, 8–9). Although the focus of the replication crisis has largely not been on physics experiments, such QRPs would be epistemically detrimental in physics as well.

In this section, I recount some of the nuances, surprises, and lessons learned from some notable history and philosophy of science research that has attended closely to the relationship between the practical and contextual details of important experiments and the epistemic significance of their results. In particular, I examine Robert Millikan's oil drop experiments to measure the fundamental electric charge, and Arthur Eddington's eclipse expedition of 1919 to test a prediction of Albert Einstein's general theory of relativity. In both cases, I focus on the epistemic justification that experimentalists may have, or fail to have, for omitting data from their final analyses.

3.1 Millikan's Orphaned Drops

In 1911 and 1913, Robert Millikan published results from his oil drop experiments on the value of the fundamental electric charge, *e*. Among his aims were

to demonstrate that all electric charges were exact multiples of a single elementary charge and to measure the value of this fundamental charge. At the time the experiments were performed, there was still debate in the physics community about whether there might be fractional charges. Thus Millikan sought to produce decisive results on that issue and to provide a more precise measurement of e than was then available.

The basic method of the experiment was to record the time it took a small drop of oil carrying an unknown number of ions to fall between two plates under the influence of gravity, and then to record the time it took the same drop to travel the same distance under the influence of a known electric field. To calculate the total electric charge on a drop, Millikan used the formula:

$$e_n = \frac{4}{3}\pi\left(\frac{9\mu}{2}\right)^{\frac{3}{2}}\left(\frac{1}{g(\sigma-\rho)}\right)\left(\frac{v_1+v_2}{F}\right)v_1^{\frac{1}{2}} \qquad (1),$$

where v_1 and v_2 are the velocities of fall and rise respectively, calculated from the measured values of the times and distances of drop travel. The other values are known: μ is the viscosity of air, g is gravitational acceleration, σ is the density of the oil, ρ is the density of air, and F is the electric field (Franklin 2016, 116).

Millikan (1911) describes the method as follows:

> The appearance of this drop is that of a brilliant star on a black background. It falls, of course, under the action of gravity, toward the lower plate; but before it reaches it, an electric field of strength between 3,000 and 8,000 volts per centimeter is created between the plates by means of the battery B, and, if the droplet had received a fractional charge of the proper size and strength as it was blown out through the atomizer, it is pulled up by this field against gravity, toward the upper plate. Before it strikes it the plates are short-circuited by means of switch S and the time required by the drop to fall under gravity the distance corresponding to the space between the cross hairs of the observing telescope is accurately determined. Then the rate at which the droplet moves up under the influence of the field is measured by timing it through the same distance when the field is on. This operation is repeated and the speeds checked an indefinite number of times. (quoted in Franklin 2016, 115)

As Franklin (2016) puts it, "Millikan's technique in manipulating the oil drops was nothing short of spectacular" (116). Millikan would stare at a single oil drop for sometimes more than four hours and took measurements for forty-seven days (116, 119).

Franklin's scholarship on the Millikan oil drop experiments is particularly interesting. Millikan was evidently "selective" regarding his data. He sometime omitted measurements that he had made on particular drops. As Franklin

recounts, in his 1911 publication on the oil drop experiments Millikan was "quite explicit about the exclusion of data" (119). In particular, he left the first four drops (smallest/slowest) and the last four drops (largest/fastest) he recorded out of the final analysis, explaining:

> These are omitted not because their introduction would change the final value of e_1, which as a matter of fact is not appreciably affected thereby, but solely because of the experimental uncertainties involved in work upon exceedingly slow or exceedingly fast drops. When the velocities are very small residual convection currents and Brownian movements introduce errors, and when they are very large the time determination becomes unreliable, so that it is scarcely legitimate to include such observations in the final mean. (quoted in Franklin 2016, 119)

However, in his subsequent paper reporting an even more precise and accurate measurement of the charge of the electron, Millikan was less transparent. Franklin (1986) tells the story as follows:

> In Millikan's famous 1913 paper, "On the Elementary Electrical Charge and the Avogadro Constant," he stated that the 58 drops under discussion had provided his entire set of data. "*It is to be remarked, too, that this is not a selected group of drops but represents all of the drops experimented upon during 60 consecutive days*." That statement is false. Millikan's laboratory notebook for that period shows that Millikan made observations from 28 October 1911 until 16 April 1912 and that he recorded data on 175 drops. Even if one were to count only observations made after 13 February 1912, the date of the first observation Millikan published, there would still be 49 excluded drops. Millikan also excluded observations within the data on a single drop and used selective calculational procedures. (229)

Insofar as Millikan ignored data he had no reason to expect were bad data except the fact that they disagreed with his preferred result, he engaged in epistemically detrimental QRPs.

Is the famed experimentalist a fraud? Franklin's conclusion is that while not an outright fraud (he gives other examples of fraud in science), Millikan can be characterized as a "trimmer" – one who leaves ostensibly relevant data out of analysis because they differ most from the mean, in order to report narrower uncertainties than if those data were included (230). Trimming data gives the appearance of a more precise measurement, a measurement with tighter error bars, than untrimmed data. Following Charles Babbage, Franklin contrasts this sort of trimming from the tails of a distribution with "cooking," meaning "selection of data to achieve agreement," for instance, "selecting only those data that will agree with a particular hypothesis or theory" (1986, 227).

Is trimming really epistemically detrimental? Does it harm the epistemology of Millikan's experiment? Trimming certainly has the potential to be epistemically detrimental. Suppose two different experimentalists seeking to measure the same physical parameter both discard outlying data in order to achieve a more precise estimate. If their reported uncertainties are artificially tight, then the two results may appear to be discordant, when they would not be had all of the relevant data been included in the analysis. This could lead to confusion and possibly even to misguided inferences about the epistemic status of the hypotheses involved in the experiments. Was Millikan's trimming in particular epistemically problematic?

Of the forty-nine drops that Millikan observed after Franklin believes that Millikan was confident in his apparatus, for twenty-two of them, Franklin's research shows that Millikan did not bother to calculate the fundamental charge. Franklin simply states: "The most plausible explanation for their exclusion is that Millikan did not need them for his determination of *e*" because he "had far more data than he needed to improve the measurement of *e* by an order of magnitude" (230). Although Franklin is not particularly worried about these twenty-two unused drops, I find it at least somewhat puzzling that Millikan would have omitted them from his analysis. As I indicated earlier in this section, taking these measurements is painstaking work requiring peering for hours through an eyepiece at pinpricks of light slowly falling and rising. Once the patience and energy had already been expended to record the data, why not use them, especially if doing so would not spoil the aim of one's experiment, as Franklin indicates was the case for these drops?

Franklin does worry, however, about twenty-seven drops for which Millikan did calculate a value of *e* and then subsequently excluded from his analysis. Of twenty-one of these, Franklin says: "Twelve were excluded because they required a second-order correction to Stokes's Law, two because of equipment malfunctions, five because they had few reliable observations, and two for no apparent reason, presumably because they were not needed" (230).

The factors motivating selectivity with regard to these twenty-one drops are not all equally compelling. It seems perfectly admissible, for instance, to reject drops because of an independently obvious equipment malfunction. Lack of sufficient reliable observations also seems like a reasonable – although conventional and thus debatable – ground for omission. The issue of corrections to Stokes's law warrants further investigation. Stokes's Law gives the drag force on a sphere (the oil drop) moving through a viscous fluid (the air) under certain conditions. In his 1911 paper, Millikan actually explores the limitations of assuming that Stokes's law strictly holds in the context of his experiments. He argues that the law "breaks down as the diameter of the sphere becomes

comparable with the mean free path of the molecules of the medium" – that is, for very small drops in Millikan's experiment (quoted in Franklin 2016, 113). He introduced a modification of Stokes's law to represent the mean free path of the molecules composing the viscous fluid, the size of the drop, and an empirically determined constant (117–121). Given that Millikan introduced this other modification to Stokes's law, why did he omit the twelve drops on account of those requiring an(other) modification? Franklin offers the following explanation: such a modification "made calculations based on their data unreliable." He explains in an endnote that he, Franklin, "made an unsuccessful attempt to make a second-order correction to Stokes's law along the lines of Millikan's first-order correction" (122, note 24). This explanation strikes me as worthy of further investigation. My comments on the twenty-two unused drops also apply to the two drops omitted "for no apparent reason."

The remaining six exclusions are even more suspect. Franklin writes that, for five of these,

> Millikan not only calculated a value of *e* but also compared it with an expected value ("1% low"). His only evident reason for rejecting these five events is that their values did not agree with his expectations. The effect of excluding the five events under discussion was to make Millikan's data appear more consistent, to make the "largest departure from the mean value . . . 0.5%." Had he included those events, the departure would have been 2 percent. (1986, 230)

Franklin suspects that Millikan trimmed his data in this way to support his reputation as a highly skilled experimenter (230). If this was Millikan's motivation, then the omission of these five drops is epistemically disingenuous in the sense that the decision was not justified on epistemic grounds. If Millikan suspected that these data had been spoiled by unfavorable experimental conditions, that would be one matter. But the fact that omitting these data would protect his reputation does not by itself furnish appropriate epistemic reasons to omit them. Such decisions should not be condoned in the epistemology of experimental physics.

We come to the final drop. Millikan had two methods for calculating *e*; one involved the total charge on a drop and the other involved changes in charge. As Franklin describes, Millikan used both methods to calculate a value for *e* from the final drop:

> That event, the second drop of 16 April 1912, was among Millikan's very best observations. It had a large number of measurements, and the two methods of calculating *e* gave results that were consistent both internally and with each other. Millikan liked it: "Publish. Fine for showing two methods . . . ." When

> Millikan calculated *e* for that event, he found a value some 40 percent lower than his other values. He dismissed the event with the comment "Won't work" and did not publish it. Millikan may have excluded that event to avoid giving Ehrenhaft ammunition in the charge-quantization controversy. This would seem to be a case of cooking. In Millikan's defense, we may note that although there are no obvious experimental difficulties with the event, the data require that not only the total charge on the drop but also each change in that charge be an integral multiple of a fractional charge, a highly unlikely event. (231)

Franklin reports that Millikan stated he exclusively used the method involving the total charge on the drop; however, in fact he used some combination of the two methods or the second method alone for nineteen out of the fifty-eight published drops, which tended to make his results seem more consistent (230). Millikan's treatment of this last drop in particular, though, seems egregious. He clearly felt his apparatus was working properly, noted "Publish," calculated the result from the data, and then abandoned the result because it failed to meet his expectations. This is just the sort of practice that "blind" or "double-blind" analyses serve to prevent – without another reason to throw out this result, this decision should be regarded as epistemically detrimental.

Franklin is not too flustered about Millikan's "selective" analyses. Franklin explains that the "effect of Millikan's trimming was quite small" and the effect of the cooking "was also small" (232). Indeed, several decades after *The Neglect of Experiment*, Franklin's appraisal is roughly the same. "The effects of both the exclusion of data and of the selective analysis of the data are negligible," and he suspects "that Millikan knew" that the effects of these questionable practices were negligible too (2016, 123). Millikan's value of *e* was 4.778×10^{10} esu with a statistical error of ± 0.002, while Franklin's reanalysis, including the published fifty-eight drops and twenty-five unpublished drops recorded after February 13, 1912, is 4.780×10^{10} esu ± 0.003 – that is, in agreement (1986, Table 8.1).

I think Franklin is too generous with Millikan. What matters is not that Millikan's trimming and cooking happened not to throw off the course of physics when viewed in retrospect. Absent clairvoyance, Millikan could not have known that in the long run his selectivity would not have derailed further research. Condoning Millikan's decision-making would equally condone QRPs today for which we do not yet have the benefit of hindsight. Surely that would be counterproductive. We can perhaps feel relief that Millikan's unjustified decisions did not happen to have majorly bad impacts on the state of our knowledge, but that in itself does not justify those decisions.

In the following section we consider another famous case in which physicists left data out of analysis and have been charged with doing so for no good reason.

This further case gives us an opportunity to inquire in more detail about the sort of reasons for omitting data that should be considered permissible.

3.2 The Sobral Astrographic Plates

Daniel Kennefick has stated that Arthur Eddington's eclipse expedition of 1919 "may well have been the most important scientific experiment of the entire twentieth century" (2019, 4). The expedition continues to be touted as the experiment that resoundingly confirmed Einstein's general theory of relativity despite decades of nuanced scholarship on the epistemology and social/historical context of the results. The near-mythological narrative is very clean: Einstein predicted that the image of a star near the limb of the sun will be displaced from its relative position when away from the sun. The apparent deflection compared to the Newtonian standard should be detectable with careful observation. In 1919, Eddington took advantage of a total solar eclipse visible from Príncipe off the west coast of Africa to make the measurement. Voila! Einstein was vindicated and an instant celebrity. General relativity took its rightful place in the canon, radically changing the landscape of physical theory forever.

There has been a truly enormous amount of scholarship on the 1919 eclipse results, their interpretation, and the epistemic fallout thereof. Analysis of this experiment covers much diverse ground, from arguing that the eclipse results were the pivotal difference-maker in the acceptance of general relativity (see also Kennefick 2009; 2012), that the public presentation of the results was significantly tilted by political motivation in the immediate aftermath of World War I (Earman and Glymour 1980), and that the results were inconsequential for the wide acceptance of general relativity among members of the physics community in comparison to the theory's success in retrodicting the advance of the perihelion of mercury, which had long vexed astronomers working in the Newtonian framework (Brush 1989; 1999). I cannot hope to offer a comprehensive discussion of this scholarship here. Instead, I will endeavor to highlight a few points of particular interest in the work on the epistemology of this experiment over the past several decades. In particular, I will focus on the question of whether Eddington was epistemically justified in omitting certain data collected during this eclipse expedition.

The persistent pedagogical narrative of this experiment papers over the significant logistical difficulties involved in a way that ultimately makes the results seem more decisive than they were at the time. The experiment caricature (just compare two photographic plates and measure the distance between dots!) obscures the expertise and finesse needed to get any useful measurements

in this context. John Earman and Clark Glymour characterized some of the fussy yet crucial details of the measurement as follows:

> In practice, the slightest mechanical change in the telescope between the taking of the two sets of photographs will alter the scale: one millimeter on the eclipse photographs will correspond to a different number of seconds of arc than will a millimeter on the comparison photographs. A displacement will occur because of the change of scale. Since the eclipse and comparison photographs are ordinarily taken months apart, one set during the day and the other at night, significant scale differences are always to be expected. In addition, small rotational and translation shifts of star images occur in the course of superposing the eclipse photograph on the comparison plate. Further, besides the displacement due to scale differences (traceable chiefly to a change in focal length of the telescope) there are displacements due to changes in the orientation of the photographic plates to the optical axis. Altogether, modern treatments of the deflection involve at least a dozen parameters (six for displacement in the direction of each of two orthogonal axes) that require the images of at least six stars for their determination. (1980, 59)

Of course, this sort of constellation of auxiliary concerns is not unique to the 1919 eclipse measurements. Nearly any interesting experiment has them. However, keeping these details in mind is crucial for understanding the epistemic significance of the results of the experiment. The eclipse expedition had two parties: with the authority of his title as Astronomer Royal, Sir Frank Watson sent Eddington and Edwin Cottingham to Príncipe and Andrew Crommelin and Charles Davidson to Sobral in Brazil. The Brazil contingent took plates with two instruments, an astrographic telescope from Greenwich that yielded nineteen plates on the day of the eclipse and another telescope with a four-inch aperture borrowed from the Royal Irish Academy (RIA) that yielded eight (73). Met with cloud cover in Príncipe, Eddington and Cottingham took sixteen plates, only two of which were usable (73). Einstein's predicted deflection was 1.74" of arc while the Newtonian value is 0.87". The original analysis of the results yielded values of 1.98" with a probable error of 0.12" (RIA telescope at Sobral), 0.86" with no reported probable error (Greenwich astrographic instrument at Sobral), and 1.61" with a probable error of 0.30" (Príncipe) (74–75). Earman and Glymour calculated the associated standard deviations in seconds of arc, yielding 0.178" (RIA telescope at Sobral), 0.48" (Greenwich astrographic instrument at Sobral), and 0.444" (Príncipe) (75). Earman and Glymour argue:

> The natural conclusion from these results is that gravity definitely affects light, and that the gravitational deflection at the limb of the sun is somewhere

> between a little below 0.87" and a little above 2.0". If one kept the data from all three instruments, the best estimate of the deflection would have to be somewhere between the Newtonian value and the Einstein value. If one kept only the results of the Sobral 4-inch instrument, the best estimate of the deflection would be 1.98", significantly above Einstein's value. (76)

Note that the results from the Greenwich instrument at Sobral might arguably be consistent with the Newtonian value. Yet the interpretation Frank Watson Dyson delivered to the Royal Society was univocal: "After careful study of the plates I am prepared to say that there can be no doubt that they confirm Einstein's prediction. A very definite result has been obtained that light is deflected in accordance with Einstein's law of gravitation" (77). William Wallace Campbell of the Lick Observatory was still unconvinced by 1923, noting that insofar as the British chose to take the Príncipe results seriously but not the results from the Greenwich instrument at Sobral, "the logic of the situation does not seem entirely clear" (78). By 1920, Eddington reported the results as 1.98" from Sobral and 1.61" from Príncipe without mentioning the results from the Greenwich instrument at all (79). The analysis offered by Earman and Glymour (1980) concluded: "The British results, taken at face value, were conflicting, and could be held to confirm Einstein's theory only if many of the measurements were ignored. Even then, the value of the deflection obtained was significantly higher than the value Einstein predicted" (50–51). Indeed, Earman and Glymour put the point quite bluntly: "Dyson and Eddington, who presented the results to the scientific world, threw out a good part of the data and ignored the discrepancies" (85). What reason is there to trust the cloud-contaminated astrographic results from Príncipe whilst ignoring the cloud-free astrographic results from Sobral except that the latter were on Newton's side? Did Dyson and Eddington engage in QRPs too?

The experiment came hot on the heels of World War I, during a time when there was a lot of skepticism, vitriol even, from scientists of nations in the Allied Powers toward those of the Central Powers, and considerable defensiveness on the part of the latter. Manifestos, public letters, condemnation, name-calling, and suggestions for boycotting members of the international scientific community circulated. Thus, an alternative explanation of Dyson and Eddington's data analysis decisions and their interpretation of the data was available. Eddington wanted peace and unity among the scientific community after the war and championed Einstein as means to that end.

At the end of the day, Earman and Glymour argued that the personalities and extensive public advocacy of Dyson and Eddington were essential to the reception of the 1919 eclipse results as confirming general relativity (52).

They describe Dyson as "a man of solid but not brilliant scientific accomplishments" who "was one of those people, familiar in every discipline, who exercise enormous personal authority well beyond the influence of their published work" (71). They conclude: "For Eddington, one of the chief benefits to be derived from the eclipse results was a rapprochement between German and British scientists and an end to talk of boycotting German science" (83). Was Eddington epistemically justified in leaving out the Sobral astrographic plates from his interpretation of the expeditions' results? Or did social considerations disrupt the epistemology of this experiment?

Harry Collins and Trevor Pinch (1993) discuss the 1919 eclipse results in their well-known book *The Golem: What Everyone Needs to Know about Science*. The overall premise of this book is that science is a golem:

> A golem is a creature of Jewish mythology. It is a humanoid made by man from clay and water, with incantations and spells. It is powerful. It grows a little more powerful every day. It will follow orders, do your work, and protect you from the ever threatening enemy. But it is clumsy and dangerous. Without control, a golem may destroy its masters with its flailing vigour. (1)

Collins and Pinch stress that all important science is controversial and empirical results are not decisive. Science, as a golem, is "not an evil creature but it is a little daft." It is "not to be blamed for its mistakes; they are our mistakes," and "powerful though it is, it is the creature of our art and craft" (2). This view puts science on a level with other human activities, which Collins and Pinch entreat us to appraise with the same ordinary logic we use in everyday contexts. They conclude:

> Scientists are neither Gods nor charlatans; they are merely experts, like every other expert on the political stage. They have, of course, their special area of expertise, the physical world, but their knowledge is no more immaculate than that of economists, health policy makers, police officers, legal advocates, weather forecasters, travel agents, car mechanics, or plumbers. The expertise that we need to deal with them is the well-developed expertise of everyday life; it is what we use when we deal with plumbers and the rest. Plumbers are not perfect – far from it – but society is not beset with anti-plumbers because being anti-plumbing is not a choice available to us. It is not a choice because the counter-choice, plumbing as immaculately conceived, is likewise not on widespread offer. (145)

I am sympathetic to this outlook in some respects. Certainly, science is a human endeavor and humans are at times magnificent craftspeople and at times bumbling fools. For Collins and Pinch, however, it is not the weight of empirical evidence that induces epistemic progress in science, it is the management of the scientific community that takes discordant empirical results and

"brings order to this chaos, transmuting the clumsy antics of the collective Golem Science into a neat and tidy scientific myth" (151). This gives too much credit to social factors. We can often identify the expert steps from the misguided stumbles. If Eddington threw out serviceable data for no good epistemic reason, we can call foul and ask better of other scientists. Collins and Pinch imply that he did. They describe Eddington's treatment of the results in the following way:

> Eddington's observations . . . were very inexact and some of them conflicted with others. When he chose which observations to count as data, and which to count as "noise," that is, when he chose which to keep and which to discard, Eddington had Einstein's prediction very much in mind. Therefore Eddington could only claim to have confirmed Einstein because he used Einstein's derivation in deciding what his observations really were, while Einstein's derivations only became accepted because Eddington's observation seemed to confirm them. Observation and prediction were linked in a circle of mutual confirmation rather than being independent of each other as we would expect according to the conventional idea of an experimental test. (45)

Collins and Pinch argue that Eddington did not have good empirical grounds to justify omitting the Sobral astrographic plates. They note that Eddington dismissed the plates as suffering from sources of systematic error, and state that if this were indeed true of the plates in question, "then Eddington would have been quite justified in treating the results as he did," but "that at the time he was unable to educe any convincing evidence to show that this was the case" and instead made "after-the-fact determinations of what the observations were taken to be" (51). Instead of a decisive test of general relativity, Collins and Pinch explain this historical episode as a profound *cultural* shift toward the broad acceptance of relativity. After Eddington's eclipse expedition, the interpretation of other previously uncertain phenomena, such as Einstein's gravitational redshift, began to fall out on Einstein's side. They use the analogy of crystal formation: "Once the seed crystal has been offered up, the crystallisation of the new scientific culture happens at breathtaking speed" (53). This spread, they suggest, was propagated by selectively omitting data: "Eddington and the Astronomer Royal did their own throwing out and ignoring of discrepancies, which in turn licensed another set of ignoring and throwing out of discrepancies, which led to conclusions about the red-shift that justified the first set of throwing out still further" (53).

Was there good reason to ignore the results derived from the plates taken at Sobral with the Greenwich astrographic instrument? Were there relevant sources of systematic error that would have justified throwing out those data for the purpose at hand? Daniel Kennefick has argued that there was.

Kennefick's interpretation, whether or not it ultimately withstands the scrutiny of further scholarship, illustrates the sort of reasons that would have been good grounds for omitting the Sobral astrographic plates from the data analysis.

The eclipse occurred on May 29, 1919. In Davidson's diary entry dated May 30, 3:00 a.m., he wrote:

> Four of the astrographic plates were developed and when dry examined. It was found that there had been a serious change of focus so that, while the stars were shown, the definition was spoilt. This change of focus can only be attributed to the unequal expansion of the mirror through the Sun's heat. The readings of the focusing scale were checked each day but were found to be unaltered at 11.0 mm. It seems doubtful whether much can be got from these plates (quoted in Kennefick 2019, 179; see also Mayo 1991, 542)

The response of the mirror to temperature posed a significant problem because changes in the scale of the plate image are difficult to distinguish from the phenomenon of interest (see Kennefick 2019, 190–192). A change in scale, just like the predicted phenomenon, would radically shift the apparent position of the stars. There is a difference: a change in scale would induce shifts greatest at the edge of a plate centered on the Sun whereas the phenomenon of interest would be greatest nearest the Sun. However, with few displaced stars to measure, these effects could unfortunately be indistinguishable in practice (191). The loss of focus of the Sobral astrographic instrument distorted the images of the stars to be measured, so that they were not circular. This was problematic because the team was trying to measure a sub-arc second effect using smushed star images that were three or four arc seconds across (192–193). Without being able to pinpoint the center of the star images, any displacement measurements taken would have been of dubious accuracy. The team attempted to determine the change in scale, but Kennefick suggests that their estimate may have been hampered by using only the right ascension coordinates of the images (193–194). Kennefick suggests that Dyson may have suspected that the team had not accurately determined the scale of these plates, and if Dyson indeed had that worry, "he was therefore justified in ignoring any result derived from that data" (195). The researchers were worried about the quality of the Sobral astrographic plates before data reduction had even begun, and thus before they would have noticed the lack of agreement of the deflection calculated from those plates with Einstein's prediction (201). Collins and Pinch are therefore wrong to think that there was a tight circle of mutual confirmation between theory and evidence here. If Kennefick is right, the reasons that Dyson and company had for omitting the Sobral astrographic plates were independent of the results generated from them.

Kennefick (2009) offers the following possible reconstruction of the experimenters' reasoning:

> If their calculation of a large change of scale in the astrographic plates was correct, then the instrument must have undergone a significant change in magnification due to the temperature change during the eclipse. That would mean that the deflection value measured was consistent with Newtonian theory. Alternatively, if one argued that the instrument might have simply lost focus, with no problematic change of scale having taken place, then the implied result was more consistent with the Einsteinian theory and with the results obtained by the Sobral 4-inch and Principe astrographic lenses. Support for the Newtonian theory was thus, in some sense, logically incompatible with the instruments having behaved in the intended manner. I suspect that line of argument strongly influenced the Greenwich team's decision to exclude the astrographic data from their final report. (42)

Insofar as the scientists involved in the 1919 eclipse expedition had reason to suspect that the usefulness of Sobral astrographic plates had been significantly degraded by the change in focus of the instrument that they noted before beginning data reduction, they had good reason to omit the deflection values calculated from those plates from the final analysis.

3.3 Lessons for Epistemology of Experiment

In this section I have discussed two well-known cases from the history and philosophy of science: Millikan's oil drop experiments and the Eddington eclipse expedition of 1919. In both cases, the researchers omitted data from analysis. In the case of Millikan, Allan Franklin has argued that while the experimenter omitted some data for no good reason, and thus may be rightfully accused of "trimming" and even "cooking" the data, ultimately these questionable practices did not harm the long-term epistemic impact of the results of the experiments. I have argued that Franklin is too generous with Millikan and that leaving out data for no good reason ought to generally be regarded as epistemically detrimental even if it happens not to derail science in a particular instance. We cannot count on the benefit of hindsight to judge whether such practices are permissible in science. Omitting data without just cause, and in particular, simply in virtue of the fact that results generated from those data disagree with the experimenter's preferred outcome, is bad practice.

In the case of the eclipse expedition, Eddington has been accused of omitting data that failed to agree with his preferred outcome, in particular omitting the estimates of the general relativistic effect of the sun on starlight near its limb as calculated from one of the three instruments used in the eclipse expeditions of 1919 in which he participated, because that calculation did not confirm the value

Einstein predicted. However, Daniel Kennefick has argued that the focus on Eddington is misguided, and when one examines the reasoning of Davidson and Dyson, one can see that they had good reason to omit the data in question from the final analysis. Far from throwing the data out based on their disagreement with Davidson and Dyson's anticipated or preferred outcome, Davidson and Dyson recognized a problem with the functioning of the instrument during the process of taking data that gave them reasonable grounds to doubt the utility of their results for the scientific purpose at hand.

The lesson displayed in both of these cases generalizes. To throw out data just because they are inconvenient is bad epistemic practice, but it can be reasonable to throw out data due to instrument malfunction. Instrument malfunction is an appropriate excuse. How can one tell that the equipment is not working properly? This question brings us back to the heart of Collins' experimenters' regress argument. If there is no way to tell whether the experimental equipment is functioning properly other than by obtaining the expected result, then there is no substantive distinction between throwing out data just because they fail to produce the anticipated result and throwing out data because the "equipment is not functioning properly." The following section explores this issue still further, by investigating the ways in which experimental physicists set about determining whether their apparatus is working appropriately for the purposes of the experiment they are trying to perform.

4 Is There an Epistemology of Experimental Physics?

Only there, in the laboratory itself, can one see how the miner sifts gold from pyrite.
–Peter Galison (1987, 19)

For many years, Allan Franklin has been developing and advocating for what he calls an "epistemology of experiment" (see e.g. Franklin 2016). His intellectual opposition in this endeavor has been sociologists of science and those working in the science and technology studies tradition, such as Harry Collins and Andrew Pickering, who emphasize the influence of social, political, and material factors on decision-making in science, including decision-making about when the results of research are credible and when some research activity ought to be concluded. Rightfully so, Franklin has resisted the extreme view that which results are deemed credible is *totally* determined by such factors. Science is a distinctive enterprise whose particular epistemic significance derives from its unique sensitivity to the way the world is, rather than the way any humans think it is or want it to be. At the same time, Franklin has, again quite rightfully, resisted the unrealistic dream of codifying the logic of science into a single universal algorithm.

Although Franklin is a physicist and historian of physics, and often discusses cases from experimental physics, his epistemology of experiment is intended to apply to experiments outside of the domain of physics as well. For instance, Franklin has a singular reverence for Gregor Mendel's experiments in plant hybridization, which he calls "The Best Experiments Ever Done!" (2016, 11). Franklin's epistemology consists of a list of strategies that scientists use to argue for the correctness of their results. He has been quite clear that this list is neither exclusive nor exhaustive. A recent version of the list is as follows:

1. Experimental checks and calibration, in which the experimental apparatus reproduces known phenomena;
2. Reproducing artifacts that are known in advance to be present;
3. Elimination of plausible sources of error and alternative explanations of the result (the Sherlock Holmes strategy);
4. Using the results themselves to argue for their validity. In this case one argues that there is no plausible malfunction of the apparatus, or background effect, that would explain the observations;
5. Using an independently well-corroborated theory of the phenomena to explain the results;
6. Using an apparatus based on a well-corroborated theory;
7. Using statistical arguments;
8. Manipulation, in which the experimenter manipulates the object under observation and predicts what they would observe if the apparatus was working properly. Observing the predicted effect strengthens belief in both the proper operation of the experimental apparatus and in the correctness of the observation;
9. The strengthening of one's belief in an observation by independent confirmation; and
10. Using "blind" analysis, a strategy for avoiding possible experimenter bias, by setting the selection criteria for "good" data independent of the final result. (4)

Some philosophers of science, even those who are not disposed to expect an algorithmic approach to scientific reasoning, may wonder if more structure is available in the composition of the strategies that Franklin highlights, and if other important strategies ought to be included in addition. For instance, Franklin's strategies 1, 3, 4, and 6 all seem to be necessary components of any credible argument for the correctness of experimental results, although 3 and 4 seem redundant and might be effectively merged. To present a credible case for the correctness of their results, experimenters must be in a position to argue that they understand the functioning of their instruments well enough

for the purposes at hand, that those instruments were well calibrated and functioning appropriately during data collection, and that plausible sources of error have been eliminated or accounted for in the presentation of the results. Without these elements, researchers would not have good reason to suppose their results to be scientifically valuable for the purposes at hand since, for all they would know, those results could have been generated from miscalibrated, malfunctioning instruments and contaminated data. Researchers claiming the correctness of results, let alone attempting to publish a peer-reviewed paper, without these elements would be roundly criticized at the first opportunity. Indeed, these elements of the case for results are often (at least in my experience in experimental physics!) extensively analyzed internally within research groups and communities. It is part of the training and culture of experimental physicists to obsessively generate and investigate alternative explanations for experimental results. One famous, and deservedly so, example of this aspect of experimental life is the case of Arno Penzias and Robert Wilson's discovery of the Cosmic Microwave Background radiation (1965). Statistical arguments – Franklin's strategy 7 – are also often necessary, depending on the subject matter at hand. They are not always necessary, however. Observing the rings of Saturn, for instance, does not require statistical arguments. Weak gravitational lensing surveys simply cannot escape them.

Three of Franklin's strategies strike me as desirable if available, but not, strictly speaking, necessary: "blind" analysis (10), or, perhaps better yet, *double*-blind analysis, independent confirmation (9), and explanation of the results using independently well-corroborated theory (5). Whether these desirable strategies are available will depend on the nature of the research at hand. Some research does not (or at least not readily) afford double-blind analysis or confirmation via independent methods. Furthermore, some extremely interesting and perfectly credible results are interesting precisely because of the lack of independently well-corroborated theory that would explain them. The initially mysterious flat galactic rotation curves published by Vera Rubin and W. Kent Ford, which would only later come to be interpreted within the paradigm of cold dark matter, are just such a case (1970).

The remaining two strategies are, I argue, optional. Reproducing known artifacts (strategy 2) is not necessary to the credibility of empirical results, although it would likely be reassuring to researchers who made the effort to do so. Of course, if a concerted effort had been made to produce known artifacts and *failed*, that would be another matter. However, for the researchers involved, such an occurrence would likely be interpreted as an indication that either the instruments involved were not functioning

appropriately and/or were not properly understood for the application, or that unaccounted-for effects were mucking up the results, thus punting the issue back to Franklin's strategies 1, 3, 4, and 6. Franklin is right to include this strategy among the litany of common and effective strategies that scientists employ in arguing for the correctness of their results, but it is not necessary to make this sort of argument in the context of every experiment. Similarly, manipulation of the object under investigation is not necessary for the delivery of credible *empirical* (not experimental in a narrow sense) results, and can in some cases even be counterproductive to studying the object of inquiry in an undisturbed state. Note that Franklin's elaboration of the manipulation strategy also ties it closely to the efforts researchers undertake to convince themselves that their instruments are functioning appropriately.

These reflections suggest the following modification of Franklin's taxonomy of argumentative strategies:

Conditions for the credible presentation of empirical results

Necessary	Properly calibrated and operating instruments, the functioning of which is adequately understood for the purpose at hand, the elimination of or accounting for plausible sources of error, and (when relevant) the use of appropriate statistical methods
Desirable	"Blind" analysis and confirmation by independent methods
Optional	Reproducing known artifacts and controlled manipulation of the object of study

An advantage of organizing Franklin's taxonomy in this way is that it renders more obvious the crucial role of activities like proper calibration in the epistemology of experiment. This is particularly appropriate, I suggest, given the important role that the topic of calibration has played in Franklin's own arguments against social constructivism in science.

In the rest of this section, I elaborate the important epistemic roles of calibration and the related, although not coextensive activity of *commissioning* in the epistemology of experiment. First, I discuss the epistemic role of calibration in the context of the debate between Franklin and Collins regarding the experimenters' regress and the roles of reasoning and social norms in experiment, invoking Eran Tal's coherentist epistemology of measurement – which purports to eschew the need for calibration *standards* to accomplish a calibration procedure – to appraise the usual story in a new light and offer a word of caution for those who pursue coherentism in this

context. I then discuss Slobodan Perović's analysis of in situ calibration, ultimately suggesting that Perović's analysis belies the significance of traditional-style calibration in his primary case study. I conclude the section by offering a preliminary exploration of the epistemic significance of the commissioning phases of experiments, a topic I suggest deserves greater attention from philosophers interested in the epistemology of experiment, and mentioning further topics for exploration that are beyond the scope of this Element.

4.1 Calibration

As Perović (2017) explains, Franklin has responded to Collins's emphasis on the experimenters' regress and its relation to what Collins characterizes as the social construction of knowledge by drawing attention to the important role of calibration procedures in arguing for empirical results. As a rough-and-ready example of calibration, which we will shortly complicate several times over, consider the process of calibrating a leak detector. A leak detector is a device used to identify and measure the magnitude of leaks in a vacuum system. Such an instrument could be, for instance, a vacuum pump equipped with a small device for detecting the presence of helium. An operator of the leak detector can then join the part or system to be tested to the inlet of the detector, evacuate it using the leak detector pump, and then carefully introduce small amounts of helium to areas on the outside of the part or system being tested to try to localize leaks. It is useful to be able to measure the severity of leaks in the system so as to ensure that the vacuum standards of the experiment in which the parts are to be deployed is met. For such measurements to be accurate, the leak detector must be calibrated. This can be done, for instance, by attaching an external leak standard, a canister filled with helium of known leak rate, probably obtained from the detector's manufacturer, to the leak detector and setting the detector's scale appropriately or running whatever more complicated calibration procedure is required. Although a leak detector is usually an auxiliary instrument to the main experimental apparatus, the same sort of procedure could apply to an instrument used for science data collection as well.

Franklin has argued that calibration disrupts vicious circularity in the experimenters' regress. The vicious regress relies on the idea that a properly functioning experimental apparatus is known to be properly functioning only insofar as the "correct" or expected result is obtained. A vicious regress makes room for non-epistemic social factors to be decisive in the interpretation of the results and the judgment of their success. However, Franklin has stressed that many experiments do not display vicious regresses of this sort. Rather, the apparatus

is established to be working properly by calibrating it using something besides the success condition of the experiment before data deemed scientifically interesting are even taken. Franklin characterizes calibration as "the use of a surrogate signal to standardize an instrument" and stresses the invocation of such standardization procedures for arguing that one's instrument is working correctly or for worrying that it is not (1997, 31).

Unfortunately, it does not seem that calibration modeled on paradigmatic cases alone can do the work of breaking vicious regresses in more exotic contexts. After describing how the scale of a new voltmeter might be calibrated by way of known voltages, Collins remarks:

> The assumption built into this procedure is that the unknown voltage acts upon the meter in the same way as the standard voltages which were applied to calibrate it. This is so slight an assumption as hardly to be worthy of the name. After all, a voltage is a voltage is a voltage! Nevertheless, it would be correct to say that during the calibration of a voltmeter, standardized voltages are used as a surrogate for as yet unmeasured signals. In more controversial science the assumptions underlying the process of calibration are of greater moment. (1992/1985, 101)

In the context of efforts to detect gravitational waves, a context that Collins returns to again and again, the novelty of the phenomenon makes "direct" calibration of the instrument impossible. Gravitational waves, "standard" or otherwise, cannot be procured from the supply cabinet like 5V batteries. In such circumstances, the surrogate nature of calibration standards seems to leave an uncomfortable inferential gap between the success of the calibration stage and the judgment that the instrument is operating properly for the purpose of measuring the previously unmeasured. In a recent conciliatory publication jointly authored by Collins and Franklin, Collins emphasizes the significance of social factors in tricky experimental contexts: "The so-called 'epistemological criteria' are necessary for establishing the existence of a new phenomenon (as Allan says) but they are not a sufficient criterion where dispute runs deep" (2016, 100).

Franklin has responded extensively to Collins's discussion of Joseph Weber's attempts to detect gravitational waves, arguing that Weber's apparatus failed the relevant calibration tests (1997, 46). That is, before even worrying about the inferential gap involved in interpreting purported detections of *novel* phenomena, Weber's instrument had an opportunity to show that it was suitable for the task ahead, and failed. Without handy gravitational wave standards, the detectors in these experiments were subjected to surrogate signals: "Scientists injected pulses of acoustic energy into the antenna and determined whether their apparatus could detect such pulses. Weber's apparatus failed to detect the

pulses, whereas each of the six experiments performed by his critics detected them with high efficiency" (44). According to Franklin, the crucial difference between Weber's experiment and his competitors' was the algorithm he used to search for signals in his data. Weber's algorithm was nonlinear while the competitors' was linear. Weber was evidently concerned that a linear algorithm would miss the gravitational wave signals. Addressing this worry, the competitors also checked to see if using the nonlinear algorithm on their own data would yield the detection that Weber claimed – it did not (45). Franklin concludes that in light of Weber's failure to properly calibrate his instrument using plausible surrogate signals considered together with several other problems with Weber's analysis, the physics community rightfully rejected his detection claims based on epistemological criteria (49).

Eran Tal's sophisticated account of calibration complicates the interpretations presented thus far. Tal argues that we are mistaken to think of calibration primarily as adjusting the readout of one instrument to reflect that of a standard, concluding that "comparison with a standard is neither necessary nor sufficient for successful calibration" (2017a, 246). If Tal's argument is correct, then it seems that we will have to revisit the responses to the threat of the experimenters' regress in order see if compelling arguments against it can nevertheless be made even with an understanding of calibration modified by Tal's insights. In order to do this, we had better understand Tal's critique of other accounts of calibration and the nature and implications of his own view.

4.1.1 Calibration and the Epistemology of Measurement

Adopting terminology from metrology, Eran Tal helpfully distinguishes between "instrument indications" and "measurement outcomes." An instrument indication is the final state of a measuring apparatus such as the position of a pointer on a dial or the display of some readout. By itself, an instrument indication is of little use for gaining information about the measurand. In order to yield such information, an instrument indication must be interpreted in light of other information about the measuring context, notably, a calibration function. On Tal's account, a measurement outcome refers to "a claim that is inferred from one or more indications along with relevant background knowledge" (2017a, 235). To make a claim about the measurand, one must make an informed inference from the relevant instrument indications, drawing on auxiliary information. The calibration of a caliper provides an illustrative example. Calipers facilitate the precise measurement of the diameter of workpieces by way of adjustable jaws that can be snugly fit around the piece whose diameter

one wants to measure. You might think that calibrating a caliper is straightforward; the scale on the caliper simply needs to be tuned so as to measure the appropriate diameter when applied to standard gauge blocks – that is, objects whose diameter has already been precisely determined (242). However, as Tal explains, a more accurate calibration of a caliper can be accomplished by "white-box calibration," which explicitly models factors that contribute to the caliper readout, "e.g. the roughness of the contact [between the caliper and the workpiece], the resolution of the readout, the temperature of the workpiece, and so on" (243). The more complicated calibration function that takes these aspects of the measurement into account will improve measurement accuracy.

Tal rightfully emphasizes that a widespread view of measurement fails to account for the epistemic significance of the work required to get a claim about a target from the state of a measuring instrument. Tal characterizes the traditional definition of calibration as "the activity of establishing a correlation between the indications of a measuring instrument and quantity values associated with a measurement standard," like establishing a correlation between the value displayed on the caliper readout and the previously established diameter of gauge blocks (243). The vertigo-inducing problem with this view is that people trying to make measurements do not have access to the true values of the measurand to use as standards for calibration. How is the diameter of the gauge block itself established? The vicious regress looms.

A delightful example of this problem of measuring standards is the now surpassed International Prototype Kilogram (IPK) (244). Between 1889 and 2019, the SI unit the kilogram was defined by reference to a specific little object, a platinum alloy cylinder carefully kept at the International Bureau of Weights and Measures in Paris. The unit of the kilogram just was, by convention, whatever the mass of that little Parisian object happened to be. When the object was used, say to calibrate another prototype mass, it was vulnerable to physical modifications. For instance, it could have been ever so slightly scuffed in the process and thereby lose mass. Even peacefully sitting around, the cylinder could gain mass by absorbing contaminants from the air, (some of) which mass the cylinder could then lose when cleaned (Girard 1990). By comparing the mass of the IPK with other prototypes, metrologists noticed relative changes in mass among the collection (Davis 2003). By definition, the mass of the IPK could not differ from exactly one kilogram, yet these cohort changes and background knowledge regarding the physical processes affecting the kilogram over time led metrologists to suspect that the actual mass of the IPK was quietly changing too.

In the face of such examples, Tal embraces a coherentist approach to calibration and the epistemology of measurement:

> Instead of comparing outcomes to true values, practicing scientists are faced with the challenge of evaluating accuracy and error by comparing measurement outcomes *to each other*. Such comparisons by their very nature cannot determine the extent of error associated with any single outcome but only overall mutual compatibility among outcomes. (2017a, 239)

On Tal's account, measurement in general is a modeling activity. To get to a measurement outcome, one has to construct a model of the measuring process such that the instrument indication relates to a representation of the measurand. Modeling measurements in this way is, according to Tal, what facilitates comparison of measurements across different specific contexts. Such comparison is essential to calibration. For Tal, "different measurement processes provide objective knowledge about the values of a quantity only once they have been *idealized in a mutually coherent and consistent manner* in terms of that quantity" (241). A measurement, then, is a prediction of the value of the measurand by way of the idealized model of the measuring process (243). Measurements are validated insofar as they mutually cohere with predictions of the purportedly same measurand made with different models of measuring processes – that is, properly informed measurement outcomes. By invoking these intermediary models of measurement processes, Tal claims to avoid the kind of operationalization that would isolate epistemologies of measurement to individual measurements (240). Calibration functions are therefore tools for *predicting* the value of the measurand from instrument indications (243).

Thus, the activity of calibration for Tal is "*the activity of modeling different processes and testing the consequences of such models for mutual compatibility*" (246). Measurements are deemed accurate insofar as the predictions from various models converge (248). On this definition, measurement "standards" have no particularly special role:

> As long as one is concerned with local, pairwise comparisons between instruments, it makes no epistemic difference whether (or which) one of the instruments is designated a "standard." The total uncertainty associated with the values being compared remains the same, and is arrived at through the same chain of inferences, regardless of such designation. The epistemic difference associated with the title "standard" appears only on a global scale, when metrologists are required to distribute uncertainties across large networks of instruments. (245)

How does Tal's model-based epistemology of measurement bear on the experimenters' regress and the hope of using calibration to break it? One could worry, for instance, that without the epistemic foundation of a measurement standard to use for calibration purposes, the regress wins. Tal's epistemology of measurement emphasizes the fact that no one has access

to the true values of the target measurands (238). This is of course correct, and it follows immediately from this fact that the true values cannot be employed for calibration purposes. Tal's view offers us instead the assurance of a coherent web of predictions, whose objectivity is supposed to be licensed by its "perspective-invariance": "the robustness of measurement outcomes across different material circumstances and representational contexts" (248–249). Recalling that measurement outcomes are inferences to or predictions of the value of the measurand, this amounts to explicating the objectivity of measurements with the convergence of predictions from a certain class of idealized models:

> Prior to their representation by an idealized model, there is no way of testing whether different instruments measure the same quantity; any agreement or disagreement among their indications may be construed as coincidental and attributed to some local feature of the instruments or environments. It is only once their idiosyncrasies are idealized away in a mutually coherent fashion that instruments can be viewed as sources of objective knowledge about a common quantity, such as temperature or frequency. (248)

Is such recourse to a coherent web of predictions from idealized models sufficient to truncate the experimenters' regress?

I am very sympathetic to Tal's epistemology of measurement and strongly endorse the desiderata that he has articulated for any epistemology of measurement. Tal argues that epistemological accounts of measurement must both accommodate the context-sensitivity of measurement and "clarify the conditions under which a measurement outcome may be justifiably deemed objective" (237). Furthermore, an epistemology of measurement must be both informed by actual scientific practice and yet retain a "critical and reflective attitude" so as to avoid falling into mere description of practices (238). These requirements are well motivated. We should ask that any epistemology of measurement stay close to practice while not losing its critical distance, and that it show how we learn from measurement in full awareness of the contextual details of actual measurement processes. While I agree with Tal that it is necessary to invoke background theorizing and information in order to wring an estimate of something like a parameter value from instrument readings, I suggest that Tal's emphasis on idealized models is misleading and that his blurring of the boundary between prediction and measurement ultimately unhelpful.

Tal emphasizes the importance of idealizing away the context-sensitive details of a measurement process in order to yield a representation of the target measurand that can be compared across contexts. This characterization makes it seem as though rarefying, or stripping away the contextual details of some

measurement is necessary to justify its objectivity. This is misleading because it is precisely in virtue of those contextual details that the measurement outcome has any epistemic utility at all. As I have argued elsewhere, the epistemic utility of any empirical result depends upon the details of its providence and this dependence is what makes comparison and amalgamation of results born out of different epistemic contexts possible (Boyd 2018b). I suggest that we subsume the epistemology of measurement under this way of thinking about a broader class of activities – generating scientifically useful results from empirical data. Empirical results, such as parameter values, only gain epistemic utility when considered in the context of a whole line of evidence stemming from the empirical data from which they have been generated, together with information about the manner in which all of the results in that line of evidence were produced: the metadata regarding the provenance of data records from which results are ultimately generated, and the metadata associated with the processing steps that transform data records into results (406–407). That recourse to such metadata is *necessary* for putting empirical results to epistemic use is evident from the fact that the collection of results considered by themselves is inconsistent, and thus cannot collectively serve as useful constraints on theorizing (409).

It may be possible to read Tal's account in alignment with the view of empirical evidence for which I advocate. He stresses, for instance, that "the context-sensitivity of measurement outcomes is a necessary precondition for the possibility of establishing their objectivity" (Tal 2017a, 237). However, I think that this point is better appreciated by conceiving of measurement outcomes as empirical results that can only be used as evidence in light of metadata about their provenance and processing, what I call "enriched evidence." It is not by stripping away the context-sensitive details that results of different measurements can be compared with one another. Rather, it is in virtue of understanding how the context-sensitive details inform data processing that scientists can reasonably compare results or use them jointly. This does involve construing the measurements within an epistemic "perspective" of sorts, which often includes modeling, but the details of provenance and processing are not lost or ignored in this process. In fact, and indeed as Tal also emphasizes, empirical results can often be improved by replacing assumptions made in data processing, or by refining aspects of data processing in light of new information, as when, for instance, prior mistakes are uncovered (237). If the context-sensitive details had been idealized *away*, this sort of revision and refinement would not be possible. Such practices are better accounted for by appreciating that empirical results cannot be divorced from the context-sensitive details of their provenance and maintain epistemic utility.

Take Tal's example of a white-box calibration procedure for a caliper. As we already saw, Tal explains how the calibration procedure makes use of a model of the measuring process composed of many modules representing different aspects of the interaction between the jaws of the caliper and the object to be measured. To get a useful measurement outcome from the indication presented by the caliper readout, it helps to include relevant details about these various aspects of the measuring process, including temperature, roughness, etc. It strikes me as misleading to construe the explicit building of these aspects of the measuring process into one's calibration function and keeping track of the specific information relevant to each application of the caliper as involving idealization and/or abstraction (cf. 237) – ignoring, for present purposes, the philosophical distinctions that can be made between idealization and abstraction. The context-sensitive information is purposefully represented and the epistemic utility of the measurement outcome would suffer from its absence.

In contrast, on the picture supplied by the enriched view of evidence, the epistemic utility of the result of the measurement depends on the assumptions baked into it. Those assumptions will inform the ways in which the result of the measurement may be responsibility deployed. Furthermore, the enriched view of evidence displays how measurement results can be revised when initial assumptions come under further scrutiny. Suppose assumptions made about the roughness of surfaces in the application of a caliper end up being revised. By keeping track of what assumptions were made in generating a measurement result, one can then judge whether the subsequent revision of those assumptions affects the epistemic utility of the result for one's purposes.

However, even those who grant that measurement outcomes are only epistemically useful when considered together with their enriching information still face the question posed earlier. Calibration has been lauded as an effective means of severing the experimenters' regress. If calibration is, as Tal argues, a web of mutually coherent measurement outcomes, can it serve that purpose?

A potential hiccup is that Tal's epistemology of measurement blurs the distinction between the activities of prediction and of measurement. On his view, scientists *predict* the value of a measurand from an instrument indication via a model of the measuring process. Tal construes this as a benefit of his view and a topic worthy of further investigation:

> Another advantage of the model-based view is that it exposes the close relationship between measurement and prediction, which has thus far remained implicit in philosophical writings about measurement. Measurement and prediction are traditionally viewed as two distinct kinds of epistemic activity, the former concerning the observation of actual states of affairs while the latter concerns the derivation of consequences from

> hypothetical assumptions. If the analysis presented here is correct, the boundaries between measurement and prediction are more permeable than previously supposed. Measurement outcomes are predictors that have been "objectified" through coherence, and measurement accuracy is a special case of predictive accuracy. These relationships between measurement and prediction suggest interesting new directions in the study of scientific evidence, as well as unexplored parallels between measurement in the natural and behavioural sciences. (2017b, 44–45)

While Tal is more careful in this statement than I am about to be, it is worthwhile to explore the consequences of the extreme suggested by his remarks in order to gauge what is at stake. If measurements were to have the same epistemic function as predictions, the experimenters' regress would pose a very serious problem. In fact, the consequences of that unity would be epistemically catastrophic well beyond the scope of the regress. Calibration was supposed to disrupt the regress because an instrument could be judged to be working correctly in virtue of something other than success at its principal aim. Support for the judgment that the gravitational wave detector is working correctly is supplied by noting its appropriate response to surrogate acoustic signals. In other words, the instrument is tested out on an input that has already been well characterized. In a measurement context, the caliper is judged to be working correctly, not merely when it yields some instrument indication in a new application, but when the caliper has been carefully calibrated using gauge blocks whose diameter has been measured by other procedures. That the calibration procedure, even for common instruments like calipers, is better achieved by modular modeling than by a linear fit between inputs and indications does not disrupt the important fact that the gauge blocks are needed to accomplish the procedure. Suppose measurements are just predictions. It would then be possible to construe successful calibration (as in, the new caliper readout returns the appropriate value when applied to independently measured gauge blocks) along the lines of Tal's view in the following way: *the predicted value of the measurand using the new model is coherent with the predicted value of the measurand using other models.* This phrasing is jarring because we typically understand predictions in the manner Tal indicates in the extended quote cited earlier: as deriving from theory or hypothesis. Coherence between predictions cannot truncate the experimenters' regress because coherence between predictions can be achieved regardless of the state of the world.

If, however, we understand measurement outcomes in the manner I suggested – as empirical results – then this awkwardness is avoided. Measurement outcomes are empirical results generated by cleverly processing data, in this case, records of instrument indications. The data are empirical in

virtue of having been produced by a causal interaction involving the worldly target. The data produced by that initial interaction are not epistemically useful themselves. They need to be cajoled into a form that is relevant to the desired information. This cajoling is achieved by processing them, often by invoking models. To put the results of such processing to epistemic use, one must consider them together with their enriching information. Theoretical and modeling assumptions are typically made in the course of data processing and interpreting empirical results, but this does not render those results predictions. Similarly, predictions can be empirically informed, but they are not themselves empirical results. Countenancing measurement outcomes as empirical results yields a picture of successful calibration that retains the essential role of the empirical. Consider an apparatus that is new or of questionable functioning. Using background knowledge about the functioning of the apparatus, the processing used to generate results from it is adjusted so as to deliver appropriate results when applied to well-characterized surrogates. The results are deemed appropriate when they are consistent with the results delivered by prior applications of other properly functioning apparatuses. Should new information bring the proper functioning of those prior apparatuses or other aspects of the background knowledge employed into question, the calibration should be adjusted accordingly.

In itself, recasting calibration in this way does not obviously resolve the issue of the experiments' regress. At this point, following Feest (2016), it is actually helpful to distinguish between two regresses, or better yet, a *circle* and a regress (35). First, there is a loop internal to an experiment, which is what Collins has called the experimenters' regress: the apparatus is judged to be working correctly when it returns the appropriate result and that the result returned by the apparatus is appropriate is judged on the basis that the apparatus is working correctly. In its most striking form: one judges that the phenomenon of interest has been successfully detected only when the detector is working properly, but one only has reason to believe that the detector is working properly when it successfully detects the phenomenon of interest. This circle can be severed by calibration: the apparatus is judged to be working correctly by returning the appropriate result when applied to independently well-characterized surrogates.

However, there is a second, or perhaps *proper* regress, which Tal's epistemology illuminates. On what basis have the surrogates been well characterized? For instance, in virtue of what are the diameter of the gauge blocks "already known" (Tal 2017a, 243)? The surrogates used in calibration are well characterized via other applications of apparatuses that have themselves been calibrated on other well-characterized surrogates, which have been well characterized via other applications of apparatuses, and so on. It is in the context

of this regress, which extends beyond the context of the focal experiment, measurement, or detection, that the possibility of a coherentist account of epistemic justification appears most promising. I cannot fully address the promise of such a view or the challenges it faces here. However, for those who would take up this line of inquiry, I offer a word of caution: a coherent web disconnected from causal contact with worldly targets will not do empirical science epistemic justice. Let me briefly say why I suspect this worry applies to Tal's epistemology of measurement, and how the view of measurement that Hasok Chang defends might be recruited to avoid it.

Tal (2016a) addresses the "problem of observational grounding" explicitly. He argues against traditional foundationalist views of measurement that confer special epistemic status to measurement outcomes in virtue of their closeness to observation by human senses. Instead, Tal argues (drawing on the work of Kent Staley) that insofar as measurement outcomes serve a special role in generating scientific evidence, it is because of their "security" (5). Measurement outcomes are more "secure" in this sense the less an epistemic agent expects them to require revision in the future. If an agent has reason to suspect that an outcome is bound to be revised in the future, that outcome is not very secure. This is certainly no foundationalism! Unfortunately, it is also a view that severs the crucial connection between the world and empirical results that any empiricist epistemology of science must retain. Tal goes so far as to state: "The epistemic credentials of measurement are not different in kind from those of other modes of quantitative estimation, such as theoretical prediction and computer simulation" (5). Without a distinction between theoretical predictions and empirical results, the evidential corpus floats free of the world of which it is supposed to inform us, stranding us with nothing but the products of our own imaginations.

Perhaps Tal's argument should give us pause: is there really any meaningful epistemic distinction to be made between theoretical predictions and empirical results? Hasok Chang's (2004) masterful study of the evolution of temperature measurements might seem at first glance to corroborate collapsing the distinction between theoretical predictions and empirical results. In particular, Chang's cases show that what is assumed, predicted, and measured can swap places as inquiry evolves. On the surface, this might look like looping with the attending threat of free-floating coherentism. But Chang argues that there is genuine epistemic progress here – rather than circles, these loops are spirals (2007, 14). Chang's view of "progressive coherentism" relies on the engine of epistemic iteration. Starting from a system of knowledge that is affirmed without foundationalist justification, by iterative refinements a type of progress is possible. Scientists have to start somewhere. In the case of temperature this start was affirming the reliability of human sensation for qualitative appraisal of

hot and cold. Thermoscopes, later affirmed in part because they largely agreed with the pronouncements of sensation, could then be used to show the limitations of sensation. Thus, on Chang's view, the starting point of inquiry is not retained unchanged. It is corrected and refined as the spiral emerges.

Moreover, despite the swapping and spiraling, there is still a meaningful distinction to be made between theoretical predictions and empirical results within Chang's progressive coherentism. Pick out a point on the spiral and one can discern the predictions from the results. Consider another case of Chang's: a series of stages in the empirical investigation of chemical analysis (14–16). On Chang's retelling, chemists first arrived at the view that chemical reactions involved the dissociation and reassociation of elements by noting the reversibility of some chemical reactions and assuming that the same held for all chemical reactions (15). Chang calls this the "component view." Employing the component view then led to discovery of some reactions in which weight was conserved, and then that in turn was assumed to hold for all chemical reactions (16). Assuming conservation of weight had significant consequences for the field:

> Most fundamentally, the focus on weight constituted an important refinement and change in the component view of chemical reactions, which had initially enabled the discovery of the conservation of weight. Weightless substances were eliminated from chemistry, even when they apparently maintained their identity through chemical combinations and decomposition. Several accepted chemical compositions were reversed. (16)

Put simply, the component view was first an empirical result, then an assumption, then a prediction that could be revised in light of new empirical results. This might look like an epistemically dubious transmutation of the theoretical and the empirical. Yet we do not need a pristine observational foundation for the epistemology of science to keep the distinction between theoretical predictions and empirical results from collapsing entirely. Certain empirical results inspired chemists to posit the component view; further empirical results obtained while working within that framework inspired them to refine it. The role that empirical results play in this progression could not be performed by theoretical predictions instead. In stating that computation can usefully be cast as a kind of measurement (2016a, 6), Tal risks losing sight of the channels through which nature constrains our theorizing. It may be useful for identifying predictions and empirical results to distinguish them by the functional roles they play in the epistemology of science at a given point in the spiral. However, it is not just anything that can play the role of an empirical result – it needs to be possible to tell a story about how the empirical result is causally downstream of the worldly target.

Chang emphasizes that coherentism as an approach to the epistemology of science problematically lends itself to relativism (2007, 4). If scientists have to start somewhere – if they have to "affirm" some knowledge system to get the whole process started – might they just as well start in different places? And might that generate equally justified but different lines of inquiry? While Chang argues for *progressive* coherentism, not methodological anarchy or relativism, he does embrace pluralism about scientific inquiry: "The point is not merely that we do not know which direction of development is right, but that there may be no such thing as the correct or even the best direction of development" (2004, 232). The pluralist aspect of Chang's view is fueled by the role he allows nonempirical virtues like simplicity, elegance, and explanatory power to play (227):

> There can be different ways of enhancing a certain epistemic virtue (e.g., explanatory power or quantitative precision in measurement) that involve belief in mutually incompatible propositions. Generally speaking, if we see the development of existing knowledge as a creative achievement, it is not so offensive that the direction of such an achievement is open to some choice. (232)

But what gives these "creative" developments any special epistemic status over other human achievements? I worry that in granting the nonempirical virtues such a strong role in determining the course of scientific developments, the epistemology of science does give way to the sort of relativism advanced by certain sociologists of science. To retain a distinctive epistemic status for scientific inquiry, the empirical adequacy must remain the deciding epistemic virtue. Empirical adequacy cannot be one virtue among many, sometimes overshadowed by, say, elegance. Subjecting theorizing to empirical constraints is essential to the scientific enterprise, and without that crucial piece, a coherentist approach to the epistemology of science will not be progressive in the right sort of way.

4.1.2 In situ Calibration

As we have seen, Franklin's response to the experimenters' regress relies on the idea that calibration is performed with respect to a surrogate signal, something that has been well characterized independently of the apparatus to be calibrated. Setting Tal and Chang's coherentist epistemologies of measurement aside for the time being, we can consider another challenge raised against Franklin's standard-centric characterization of calibration. Presenting a case study involving the measurement of the mass of the top quark (Mt) at the Large Hadron Collider (LHC), Perović argues that in characterizing the calibration process as typically

independent of the focal phenomena of the intended research, Franklin's view of calibration is too narrow. In particular, Perović recounts how in situ calibrations in LHC experiments iteratively feed measurements of the parameters of interest into the calibration of the overall instrument in order to continuously improve measurements of those very parameters. Perović admits:

> It would be circular, of course, to use the Mt reconstruction from the current measurement and average it with the Mt value that is used as a constraint in the calibration. Rather, the procedure is analogous to a very long string of steadily improving measuring apparata where each new apparatus uses the best results of the previous one for the calibration relying on various constraints and parameters. (2017, 326)

In light of his analysis of in situ calibration at the LHC, Perović defines calibration as "*any combination of experimental techniques that ensures the proper functioning of the apparatus based on already-known phenomena*" (317). He argues that his case study demonstrates that "Franklin's view that calibration of the apparatus does not depend on the outcome of the experiment ... requires thorough rethinking" (327). In particular, Perović argues that in situ calibration of this sort intertwines elements of measurement and apparatus validation in such a way that calibration is not totally independent of the measurement of primary scientific interest (328). Despite this nuance, Perović also argues that the social constructivist is not thereby straightforwardly vindicated. Even in the case of in situ calibration, "there are theoretical reasons, as well as technical reasons – concerning particular processes occurring in the apparatus – that justify calibration" (328). That is, Perović concludes that we are not forced to appeal to social factors in order to explain how the researchers arrive at agreement that their instrument is functioning properly and that the results thereby obtained are of serious scientific interest. In "relaxing" the characterization of calibration to allow for "entanglement" between calibration and measurement, Perović insists that we need not necessarily fall into the dire straits of the experimenters' regress (330).

I heartily endorse the methodological approach Perović advocates. The epistemology of experiment is more fruitfully approached by close examination of science in practice than arguing about conceptual distinctions divorced from real methods (Perović 2017, fn 14). Having the details of the in situ calibration case on the table is valuable, and I appreciate that Perović's argument threads between an overly dogmatic empiricist approach that would ignore actual scientific methodologies in order to defend the epistemic integrity of scientific reasoning practice-be-damned, and also an overly pessimistic view of the reasons, complicated though they may be, that scientists provide for the

integrity of their conclusions. However, I suggest that Perović's in situ calibration case harbors a particularly important role for independent calibration of the sort that Franklin emphasizes, which Perović's analysis downplays.

In particular, the iterative process of refining the parameter values used in LHC experiments that Perović details is initiated by appeal to the results of an independent experimental instrument – the Tevatron (a prior-generation particle accelerator at Fermilab that operated between 1985 and 2011):

> In order to calibrate the LHC and thus validate the process of determining whether a novel phenomenon at a particular energy is an artifact, noise, or a genuine phenomenon (e.g. whether the expected signature of the Higgs boson are genuine), past data concerning a well-known phenomenon such as the top quark are used as calibration values. (320)

As Perović explains, "initially, during the *commissioning phase*, are Tevatron data alone used for the calibration," and this procedure "matches the calibration procedures Franklin focuses upon" (322). He reiterates later that the commissioning phase "looks pretty much like standard calibrating procedures described by Franklin" (327).

> The case suggests to us at very least that the idea that the calibration of the apparatus does not depend on the outcome of the experiment should be accepted only very cautiously and conditionally. The dynamics of the calibration in the LHC case is such that the point is valued without crucial caveats only in the commissioning phase. (328).

In other words, the initial calibration procedure involved checking LHC results against Tevatron results for the same parameter values. Experimental apparatuses at the LHC were calibrated, at least initially, with reference to trusted values from the Tevatron. The more bootstrap-like procedures of in situ calibration commenced after this initial check against an independent instrument. Would physicists working with LHC data have been as confident that their machine was operating properly without these initial calibration procedures? I doubt that the usefulness of in situ calibration as Perović describes it could have been motivated without these initial procedures performed in the commissioning phase. It seems to me that a Franklin-style focus on independent calibration procedures in the epistemology of experiment is actually vindicated by Perović's case: without data like that from the Tevatron, the in situ calibration could not have gotten off the ground in the first place.

In the final part of this section, I turn to a topic that I believe has not received due attention in philosophy of science. We saw Perović mention the "commissioning phase" of LHC operation; however, he does not provide a characterization of that aspect of experimental research as such or explicate

its relationship to the epistemology of experiment. I suggest it may be fruitful to consider the epistemic role of commissioning in its own right.

4.2 Commissioning

Empirical research often relies upon apparatuses that take some serious time and attention to get up and running for their proper use in science. Speaking broadly, we can break the preparation for an experiment into a sequence of phases: design, construction, and commissioning (which may include "engineering runs" and calibration). Depending on the nature of the experiment, there may be other preparatory phases too such as prototyping and simulation. The commissioning phase in particular serves as a basis for the epistemic significance that the researchers accord to the results of the experiment. In this section I offer a preliminary discussion of commissioning and its epistemic significance in the epistemology of experiment, using the KATRIN experiment as an illustrative case.

The term "commissioning" is also used in primarily engineering contexts, such as the operation of nuclear power plants. As one article explains: "The results of commissioning have to demonstrate that the requirements and intentions of the design and the intentions of the designers, as stated in the safety analysis report, have been met and that the unit is ready for a long-lasting and successful operational phase" (Grauf 2012). The role of a commissioning phase in research contexts is not dissimilar. This is no wonder, since preparing an apparatus for science involves engineering work. For example, Richard Hills, the project scientist for the massive Chilean radio array ALMA, stated in a presentation to the ALMA Science Advisory Committee that, quite simply, the point of their commissioning efforts would be to "Make the system into a telescope – one capable of making the specified astronomical observations" (2009, slide 6). For ALMA, the commissioning phase involved design tests and debugging the electronics, antennas, infrastructure (like power), and software (Hills 2009).

To take another example, the publication reporting the results of the KArlsruhe TRItium Neutrino (KATRIN) experiment collaboration's commissioning of the vacuum system of their main apparatus frames the success of that phase of work as follows: "The vacuum system has to maintain a pressure in the 10^{-11} mbar range. It is demonstrated that the performance of the system is already close to these stringent functional requirements for the KATRIN experiment" (Arenz, Babutzka, Bahr, et al. 2016, abstract). For KATRIN, commissioning the vacuum system of the main spectrometer included, for instance, baking everything that could be baked to promote outgassing from the metals

used in the vacuum systems and finding and fixing leaks in the system that would prevent achieving the vacuum needed for the desired experiment.

Roughly, then, in scientific contexts, commissioning refers to the *preparation of an apparatus for routine performance according to the aims of the research context*. These days, at least for larger collaborative experiments, these research aims and the performance requirements they imply are often explicitly stated and published in design articles prior to beginning the construction of the experiment. Indeed, significant research is needed to specify these research aims and requirements in the first place. While the successes of the commissioning phase of an experiment might seem primarily pragmatic – for example, an ultra-high vacuum was achieved – these pragmatic wins have epistemic import, as I aim to make clear in the following discussion.

Generally speaking, there are two particularly important benchmarks in a commissioning phase. The first is when the apparatus is sufficiently assembled and functioning for basic operation. In the context of a telescope this is often referred to as "first light," when the telescope is sufficiently a telescope to make some astronomical image. That image may be quite far from being of any serious scientific interest, due to much required further work. Similarly, in accelerator physics like that conducted at the LHC, the "first beam" of an accelerator is celebrated when a beam can be successfully produced and steered through the instrument. For KATRIN, it was important to demonstrate the successful transport of electrons from their tritium source (after practicing with non-tritium-sourced electrons and ions) through to the detector end of the apparatus.

A second important phase marks the transition from taking "engineering data" to "science data." After the primary function of the instrument has been demonstrated ("we can see *something*!") operations still need to be developed and checked to have a hope of satisfying the aims of the experiment. Like other aspects of experimental methods, the precise nature of the procedures used to transition an instrument from basically functional to the realization of its full science capabilities will vary with context and are worth investigating in detailed case studies.

As a brief illustration, consider the efforts of the KATRIN collaboration in the period between when they first injected the system with tritium and when they were confident enough to publish their first major science result improving the upper limit on the neutrino mass. The purpose of the KATRIN experiment is to estimate the mass of the electron antineutrino by measuring the shape of the end point of the energy spectrum of beta decay electrons from tritium. Solar neutrino experiments demonstrating that neutrinos oscillate between their

three flavors imply that neutrinos are not massless as stipulated by the Standard Model of particle physics. Tritium beta decay yields helium-3, an electron, and an electron antineutrino. By studying the energy spectrum of the electrons produced by these decays, the KATRIN collaboration looks for a small distortion in the tail of the spectrum where the nonzero mass of the election antineutrino should nibble away at energy that otherwise would have been imparted to the electron. The KATRIN experiments aims to measure this neutrino mass with a sensitivity of 0.2 eV. Several significant technical challenges to meeting that goal were identified in the design stage of the experiment:

1. Long-term recirculation and purification of tritium on the kCi scale,
2. 10^{-3} temperature stability at 27 K,
3. Extreme high vacuum ($<10^{-11}$ mbar) at very large volumes ($\approx$ 1,400 m^3),
4. Large number of superconducting magnets ($\approx$ 30),
5. Ppm stability for voltages in the 20kV range, aim to reach ppm absolute precision as well,
6. Simulations and Monte Carlo studies (Angrik, Armbrust, Beglarian, et al. 2005, 53)

Any one of these elements is ambitious enough to make a seasoned experimenter nervous. While working as an engineer at one of the laboratories involved in KATRIN, I was told that one collaboration member, overwhelmed by the variety of catastrophic failure modes of the experiment, referred to it as something like "the flying purple unicorn." A single superconducting magnet can be temperamental and dangerous. Add to that "challenging" high voltage, high vacuum, and cryogenic requirements, plus a windowless gaseous (i.e. scary) highly pure tritium source, and you have a precarious situation. Careful commissioning to test out these diverse and difficult specs of the experimental apparatus were thus essential to ensuring that the desired experiment could run.

The 70-meter-long KATRIN apparatus includes the windowless gaseous tritium source from which the tritium decays, a transport and pumping section through which the beta decay electrons are guided by superconducting magnets to the main spectrometer. The main spectrometer vessel is a stainless steel tank of 1,400 m^3 in volume, weighing approximately 200 tons, which is supposed to maintain an ultra-high vacuum. This section of the instrument was constructed in Deggendorf and is so large that it could not be transported directly overland to Karlsruhe. The collaboration explained. "There is a slight problem of transportability from Deggendorf to Karlsruhe: The tank is too big for motorways, and the canal between the rivers Rhine and Danube has to be ruled out too. Thus instead of a journey of about 400 km, the spectrometer has to travel nearly 9000 km" down the

Danube to the Black Sea, through the Aegean and Mediterranean Seas through the Strait of Gibraltar, up through the Atlantic to the Rhine from the other end (www.katrin.kit.edu/213.php). There is an outstanding picture from this expedition of the hulking vessel of the main spectrometer escorted by police cars through Leopoldshafen to the delight of a packed crowd of onlookers, at the moment where it is barely scraping between the roofs of two houses, looking like an embarrassed grounded blimp. An enlarged and framed version of this photograph greeted me every day as I walked through the front door to the lab when I worked at CENPA, the feat proudly displayed on the wall next to our mailboxes. Just think: Boyle thought he had problems with leaks!

The mass spectrometer transports only electrons whose kinetic energy is above a certain threshold, which can be altered by the experimenters. Finally, at the business end of the mass spectrometer, there is a focal plane detector that counts the electrons. With the beamline components in place, in 2016, the collaboration tested the alignment of the magnets, demonstrating that the instrument could transport electrons and ions and also block positive ions (Aker, Altenmüller, Arenz, et al. 2020, 3). The following year, they tested spectroscopic performance with a ^{83m}Kr source and checked the calibration of their high voltage system to the parts-per-million level (ibid.). Satisfied with these preliminary commissioning procedures, the collaboration was ready to introduce tritium to the system. In the "First Tritium campaign" (FT campaign), the source was limited to 0.5% of operational activity by mixing the tritium with pure deuterium as a safety precaution. For the collaboration, "[a] key aspect of the FT campaign was to demonstrate a source stability at the 0.1% level on the time scale of hours" (ibid.) They were able to demonstrate time stability of key parameters affecting the stability of the source within design specifications over 12 days — success (ibid.). Other indicators of source stability were also checked. For instance, the rate of electrons that make it through to the detector also indicates source stability, and this rate "was demonstrated to be stable on the 0.1% level over a duration of 5 h" (ibid., 4). The tritium commissioning campaign had other objectives as well:

> Beyond these successful stability measurements, a major goal of the FT campaign was to record tritium β-electron spectra. The objectives of these spectral measurements were (1) to compare various analysis strategies, (2) to test the spectrum calculation software, and 3) to demonstrate the stability of the fit parameters in the analyses. (ibid)

They tested 30 different thresholds for the electron kinetic energies, in a larger range that would be employed in the actual measurement procedure in order to obtain significant statistics despite the reduced source activity, "gain confidence in our calculation of the spectrum over a wider interval" and

"perform a search for sterile neutrinos in the 200 – 1000 eV mass range, which is the subject of a separate publication" (ibid. 5). Based on prior research efforts to measure the neutrino mass, the collaboration was justifiably concerned about the sensitivity of the spectral shape to unknown systematics: "The analysis heavily relies on a precise description of the spectral shape including all relevant systematic effects and a robust treatment of systematic uncertainties. Any unaccounted-for effect and uncertainty can lead to systematic shifts of the deduced neutrino mass" (ibid.) For reassurance, the collaboration assigned two teams to carry out the analysis independently using different calculations and software. When the teams returned results that agreed within 4%, the collaboration reported that gave them "high confidence in our analysis tools" (ibid.) This analysis used only 82 of the 116 "scans" (a full sequence of all of the energy thresholds) recorded during the FT campaign (ibid., 8, although curiously page 5 gives the total scan number as 122 — I have asked a KATRIN collaboration member about this who said she would inquire about it internally, and have as yet to hear the explanation). The collaboration refers to this subset as the "golden" data set. They account for the omitted scans as follows:

> (1) 27 scans were performed at a different [source] column density for testing purposes and are analyzed separately, (2) we exclude four scans where different [high voltage] setpoints were shown . . . (3) we exclude the last two scans and the first scan, as the [deuterium-tritium] concentration dropped by several percent. (ibid)

The collaboration investigated various sources of systematic error including the column density, tritium concentration, the probability of endpoint electrons losing energy due to inelastic scattering, magnetic fields of the approximately 60 magnets included in the instrument, electric potentials in the source and spectrometer, rotational and vibrational states of the molecular tritium used in the source, and the efficiency of the detector, explaining how these were monitored and the extent to which the variations recorded would be tolerable in experimental conditions. The collaboration observed a 350 mcps background rate, which they hoped to reduce to under 100 mcps with further measurements and refinements (ibid., 12). They were able to demonstrate that the final results of their analyses were independent of the source column density and of the scanning mode (increasing voltage vs. decreasing voltage vs. random mode), the energy range of data used in the fit, and the spatial distribution of the pixels in the detector (ibid., 15). The collaboration concluded: "All these properties are essential prerequisites for the neutrino mass measurements" (ibid., 16). With this work accomplished, the collaboration looked forward to increasing the source activity to the nominal operating value.

Of particular interest in relation to the experimenters' regress, is the fact that the FT campaign measurements, performed during the commissioning phase, did not seek to produce the phenomenon of ultimate interest. In particular, the experimenters did not attempt to set a limit on the neutrino mass in the FT campaign. Because the FT campaign used a source of significantly reduced activity compared to nominal operating specifications, measurements made during the campaign would only have been sensitive to neutrino masses at around 6 eV, which was significantly larger than the best mass measurements that had already made by other experiments at the level of 2 eV (ibid., 4). In fact, for the purposes of the FT campaign, the experimenters set the neutrino mass parameter to zero in their analysis and instead used measurements of value of the spectrum endpoint "as a proxy to evaluate the analysis results" (ibid.) Rather that aiming to demonstrate the instrument's capacity to successfully measure the very phenomenon of interest in the experiment to come, the function of these measurements during the commissioning phase was to try out and to compare various possible analysis approaches, test the relevant software and run other validation checks that would ultimately be useful in arguing for their experimental results down the line. One could object to that, like the experiment ultimately desired, the campaign recorded the end point of the beta decay electron energy spectrum thus suggesting that "success" of the commissioning campaign depended on the detection of the very phenomenon of interest in the experiment proper thereby instantiating Collins' regress. However, this interpretation would miss the fact that success of the experiment would ultimately be achieved at 0.2 eV sensitivity.

In their publication reporting the KATRIN collaboration's first measurement of the upper limit on the neutrino mass, they present results from four weeks of data-taking (Aker, Altenmüller, Arenz, et al. 2019). By having "commissioned the entire setup by a series of dedicated measurements" over the course of several years, which "demonstrated that all specifications are met, or even surpassed by up to 1 order of magnitude, except for the background rate", the collaboration was willing to take science data (ibid., 4). The over-spec background rate had been studied in detail and attributed to decay products of ^{210}Pb implanted on the inner surface of the spectrometer during construction from exposure to ambient air (ibid.). Another source of higher-than-anticipated background events is ^{219}Rn atoms from certain vacuum pumps attached to the main spectrometer. The collaboration attempted to remedy this problem by installing special baffles at the inlets of the pumps, but a layer of H_2O covering the inner surface of the main spectrometer "originating from an imperfect bake out of the prespectrometer" introduced a layer of H_2O on the baffles themselves, disrupting their ability to trap the offending ^{219}Rn (ibid., 5).

Due to problematic drifts in the source column density, attributed to "radiochemical reactions of T_2 with the previously unexposed inner metal surface of the injection capillary", the experiment ran at a column density a factor of 5 below the nominal operating value (ibid., 5).

Recall Collins' experimenters' regress. Introducing it in his book *Changing Order* in the context of his extended discussion of early efforts at constructing gravitational wave detectors, Collins characterizes the regress as follows:

> What the correct outcome is depends upon whether there are gravity waves hitting the Earth in detectable fluxes. To find this out we must build a good gravity wave detector and have a look. But we won't know if we have built a good detector until we have tried it and obtained the correct outcome! But we don't know what the correct outcome is until . . . and so on *ad infinitum*. (1992/1985, 84)

Explicating the threat of regress in this way portrays the experimenters' efforts to check that the instrument is working as needed to perform the experiment of interest and the experiment itself as identical. This is deeply misleading because experimenters do a lot of work to assure themselves that their experimental apparatus is ready for science applications before they are willing to record data that they will take seriously with respect to the primary aims of the experiment. Calibration, often lots of it, is generally a part of what happens during the commissioning phase of an experiment, when this preparatory work is accomplished. But if Perović is correct in his analysis of in-situ calibration at the LHC, calibration of a certain nature may extend far beyond the initial preparatory stages of an experiment, perhaps even becoming intimately intertwined with the measurement process. Even in a much more routine sense, calibration procedures may feature in the experiment throughout its operational phase. In KATRIN, for instance, the rear end of the tritium source section of the beamline is capped by an electron gun for use as a calibration source as needed (Aker, Altenmüller, Arenz, et al. 2020). Calibration is part of what happens during commissioning, but neither exhausts the activities of commissioning nor is limited to the confines of that phase of the experiment.

As can be seen from the example of KATRIN's FT campaign, which was only one dramatic part of the longer commissioning phase, the tests, checks, and troubleshooting that occur during commissioning are important aspects of the epistemic support the experimenters offer for their ultimate results. In this phase, the experimenters, engineers, and technicians attempt to demonstrate that the instrument "as built" can in fact perform to design specifications demanded by the science goals. They try out different experimental strategies and finalize decisions about how to run the intended experiment. They uncover

unanticipated difficulties (like higher than desired background rates) and do what they can to understand and remedy them so that the experiment can go on. The critical work accomplished in this phase furnishes some of the important arguments that the researchers need in order to justify their ultimate interpretations of the results of the experiment. Philosophers interested in the epistemology of experiment would evidently do well to explore the methods and arguments employed in commissioning phases of experiments, in detailed case studies.

Explicitly naming a phase of an experiment "commissioning" is a widespread practice today. For large, complicated, expensive, and technically difficult experiments, such a phase is particularly prudent. Experimenters are often wise to try out part of their instrument (e.g. a small fraction of the antennae that will ultimately form a large array), their instrument at reduced power (as when accelerators initially run at diminished energies compared to that of which they are capable at full operation), or with surrogates that are less dangerous or finicky (such as relevantly similar radioactive isotopes of reduced activity). However, it also seems plausible to look for something functionally serving as a "commissioning phase" in many experimental contexts from the past or at smaller scales. Franklin forgives Millikan for not including the oil drop measurements he made in his trials between October 28, 1911, and February 13, 1912, in his final calculations of the fundamental charge published in 1913, because Millikan was not yet confident in his apparatus (1986, 230). This seems appropriate. It is reasonable to begin to use and troubleshoot one's instrument before recording science-worthy data. Indeed, the successes or failures of this preparatory phase of the experimental work is an important part of what eventually makes the results of the experiment compelling or not.

Although these issues deserve further attention, my aim has been to provide some preliminary reasons in support of the idea that grounds for some of the arguments that scientists *must* make in order to argue that their experimental apparatus works to their satisfaction – arguments that their instruments are properly calibrated and operating, that the functioning of those instruments is adequately understood for the purpose at hand, that plausible sources of significant error have been eliminated or accounted for, and that the data processing to be used is appropriate for the desired application – are often found in calibration and commissioning activities. The latter in particular deserve further philosophical attention.

4.3 Further

My aim in this Element has been to introduce you to some of the main themes of the scholarship thus far on the epistemology of experimental physics and to offer some reflective commentary. I have chosen to recount some of the cases and arguments that stood out to me in particular as I was introduced to the epistemology of experiment in physics, and to focus on *auxiliaries*, *regress*, and *excuses*. Admittedly, this is a conservative approach. Many worthy topics and perspectives have been left out. While Bacon's observations of hot horse shit and the odd contents of Boyle's larder are granted several paragraphs, I do not attempt a discussion of Ibn al-Haytham's much earlier sophisticated investigation of optics or epistemology. I do not discuss the extent to which there are interesting differences between the epistemology of experiment in physics and the epistemology of experiment generally, or in other specific scientific disciplines such as biology. The rise of Bayesian approaches, machine learning, and statistical methods broadly speaking in experimental physics deserve further attention. So do issues surrounding noise, artifacts, systematics, uncertainties, and error. Recent work on the social epistemology of experimental physics, for instance on collaborations and incentive structures, is also absent. The literature on science and values is mostly passed over. Conflict between epistemic priorities and moral imperatives – I think particularly of the reckoning necessary in the astronomy community with regard to relationships among Indigenous peoples and mountains with favorable seeing – deserve thorough and lucid treatment elsewhere.

For those who will continue to explore these topics, the cases and arguments of this Element offer the following advice. Key decisions made in empirical research — the judgment that the instrument is working properly, for instance, or that certain data should be omitted from analysis — require epistemic justification. The fact that such justification is required as a general feature of empirical research can therefore be represented in the epistemology of experiment, perhaps along with some mid-level strategies of the sort that occur in commissioning such as using well-characterized test signals and demonstrating that required design benchmarks have been achieved concretely. Yet the reasons that can appropriately serve as justification for key decisions in empirical research are often furnished by particular contextual details. Philosophical investigation of the epistemology of experimental physics thus needs to attend to such details if it aims to deliver normative arguments regarding science in practice.

References

Ackermann, R. J. (1985). *Data, Instruments, and Theory*. Princeton University Press.

Aker, M., K. Altenmüller, M. Arenz et al. (2019). Improved Upper Limit on the Neutrino Mass from a Direct Kinematic Method by KATRIN. *Physical Review Letters*, 123, 221802.

Aker, M., K. Altenmüller, M. Arenz et al. (2020). First Operation of the KATRIN Experiment with Tritium. European Physical Journal C, 80, 264.

Angrik, J., T. Armbrust, A. Beglarian et al. (2005). KATRIN Design Report 2004, FZKA Scientific Report 7090. Germany.

Anstey, P. R. (2014). Philosophy of Experiment in Early Modern England: The Case of Bacon, Boyle and Hooke. *Early Science and Medicine*, 19, 103–132.

Arenz, M., M. Babutzka, M. Bahr et al. (2016). Commissioning of the vacuum system of the KATRIN Main Spectrometer. *Journal of Instrumentation*, 11, P04011.

Bacon, F. (2000/1620). *The New Organon*. Edited by L. Jardine and M. Silverthorne. Cambridge University Press.

Boyd, N. M. (2017). Franklin's Field Guide to Scientific Experiments. *Philosophy of Science*, 84, 586–594.

Boyd, N. M. (2018a). "Scientific Progress at the Boundaries of Experience". Doctoral Dissertation, University of Pittsburgh. http://d-scholarship.pitt.edu/id/eprint/33843.

Boyd, N. M. (2018b). Evidence Enriched. *Philosophy of Science*, 85, 403–421.

Boyle, R. (1999/1661). Two Essays, Concerning the Unsuccessfulness of Experiments. In Hunter, M., and E. B. Davis, eds., *The Works of Robert Boyle, Vol. 2: The Sceptical in Chymist and Other Publications of 1661*. Oxford Scholarly Editions Online, pp. 35–82. https://www.oxfordscholarlyeditions.com/view/10.1093/actrade/9781138764699.book.1/actrade-9781138764699-div1-10?r-1=1.000&wm-1=1&t-1=contents-tab&p1-1=1&w1-1=1.000

Boyle, R. (1672). Some Observations about Shining Flesh, Made by the Honourable Robert Boyle; Febr. 15. 1671/72. And by Way of Letter Addressed to the Publisher, and Presented to the R. Society. *Philosophical Transactions (1665–1678)*, 7, 5108–5116.

Brush, S. G. (1989). Prediction and Theory Evaluation: The Case of Light Bending. *Science*, 246, 1124–1129.

Brush, S. G. (1999). Why Was Relativity Accepted? *Physics in Perspective*, 1, 184–214.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Chang, H. (2007). Scientific Progress: Beyond Foundationalism and Coherentism. *Royal Institute of Philosophy Supplement*, 61, 1–20.

Collins, H. (1992/1985). *Changing Order: Replication and Induction in Scientific Practice*. University of Chicago Press.

Collins, H., and T. Pinch. (1993). *The Golem: What Everyone Needs to Know about Science*. Cambridge University Press.

Daston, L., and P. Galison. (2007). *Objectivity*. Zone Books.

Davis, R. (2003). The SI Unit of Mass. *Metrologia*, 40, 299–305.

Duhem, P. (1991/1954). *The Aim and Structure of Physical Theory*. Princeton University Press.

Dumitru, C. (2013). Crucial Instances and Crucial Experiments in Bacon, Boyle, and Hooke. *Society and Politics*, 7(1), 45–61.

Earman, J., and C. Glymour. (1980). Relativity and Eclipses: The British Eclipse Expeditions of 1919 and Their Predecessors. *Historical Studies in the Physical Sciences*, 11(1), 49–85.

Feest, U. (2016). The Experimenters' Regress Reconsidered: Replication, Tacit Knowledge, and the Dynamics of Knowledge Generation. *Studies in History and Philosophy of Scienc*e, 58, 34–45.

Franklin, A. (1986). *The Neglect of Experiment*. Cambridge University Press.

Franklin, A. (1993). *The Rise and Fall of the Fifth Force: Discover, Pursuit, and Justification in Modern Physics*. American Institute of Physics.

Franklin, A. (1997). Calibration. *Perspectives on Science*, 5(1), 31–80.

Franklin, A. (2002). *Selectivity and Discord: Two Problems of Experiment*. University of Pittsburgh Press.

Franklin, A. (2013). *Shifting Standards: Experiments in Particle Physics in the Twentieth Century*. University of Pittsburgh Press.

Franklin, A. (2016). *What Makes a Good Experiment? Reasons and Roles in Science*. University of Pittsburgh Press.

Franklin, A., and H. M. Collins. (2016). Two Kinds of Case Study and a New Agreement. In Sauer, T., and R. Scholl, eds., *The Philosophy of Historical Case Studies*. Springer, pp. 95–121.

Franklin, A., and S. Perović. (2019). Experiment in Physics. *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/entries/physics-experiment

Fraser, H., T. Parker, S. Nakagawa, A. Barnett, and F. Fidler (2018). Questionable Research Practices in Ecology and Evolution. *PLoS ONE*, 13 (7), e0200303.

Galison, P. (1987). *How Experiments End*. University of Chicago Press.

Galison, P. (1995). Context and Constraints. In Buchwald, J. Z., ed., *Scientific Practice: Theories and Stories of Doing Physics*. University of Chicago Press, pp. 13–41.

Galison, P. (1997). *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press.

Girard, G. (1990). *The Washing and Cleaning of Kilogram Prototypes at the BIPM, Bureau International Des Poids Et Mesures*.

Grauf, E. (2012). Commissioning of Nuclear Power Plants (NPPs). In Alonso, A., ed., *Infrastructure and Methodologies for the Justification of Nuclear Power Programmes*. Woodhead, pp. 741–772.

Hacking, I. (1983). *Representing and Intervening*. Cambridge University Press.

Hacking, I. (1989). Extragalactic Reality: The Case of Gravitational Lensing. *Philosophy of Science*, 56(4), 555–581.

Hills, R. (2009). ALMA Commissioning and Scientific Verification: Summary and Status. ALMA Scientific Advisory Committee (ASAC) Face-to-Face Presentation. https://safe.nrao.edu/wiki/pub/ALMA/ASAC13Oct09Agenda/Commissioning_ASAC.pdf

Holton, G. (1969). Einstein, Michelson, and the "Crucial" Experiment. *Isis*, 60 (2), 132–197.

Kennefick, D. (2009). Testing Relativity from the 1919 Eclipse: A Question of Bias. *Physics Today*, 62(3), 37–42.

Kennefick, D. (2012). Not Only Because of Theory: Dyson, Eddington, and the Competing Myths of the 1919 Eclipse Expedition. In Lehner, C., J. Renn, and M. Schemmel, eds., *Einstein and the Changing Worldview of Physics. Einstein Studies, vol 12*. Birkhäuser Boston, pp. 201–232.

Kennefick, D. (2019). *No Shadow of a Doubt: The 1919 Eclipse That Confirmed Einstein's Theory of Relativity*. Princeton University Press.

Kheirandish, E. (2009). Footprints of "Experiment" in Early Arabic Optics. *Early Science and Medicine*, 14, 79–104.

Mayo, D. (1991). Novel Evidence and Severe Tests. *Philosophy of Science*, 58 (4), 523–552.

Nauenberg, M. (2015). Solution to the Long-Standing Puzzle of Huygens' "Anomalous Suspension." *Archive for History of Exact Sciences*, 69(3), 327–341.

Penzias, A. A., and R. W. Wilson. (1965). A Measurement of Excess Antenna Temperature at 4080 Mc/s. *Astrophysical Journal*, 142, 419–421.

Perović, S. (2017). Experimenters' Regress Argument, Empiricism, and the Calibration of the Large Hadron Collider. *Synthese*, 194, 313–332.

Rubin, V. C., and W. K. Ford Jr. (1970). Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. *Astrophysical Journal*, 159, 379–403.

Sargent, R. (1994). Learning from Experience: Boyle's Construction of an Experimental Philosophy. In Hunter, M., ed., *Robert Boyle Reconsidered*. Cambridge University Press, pp. 57–78.

Sargent, R. (1995). *The Diffident Naturalist: Robert Boyle and the Philosophy of Experiment*. University of Chicago Press.

Shapin, S., and S. Shaffer. (2011/1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press.

Tal, E. (2016a). How Does Measuring Generate Evidence? The Problem of Observational Grounding. *Journal of Physics: Conference Series*, 772, 012001.

Tal, E. (2016b). Making Time: A Study in the Epistemology of Measurement. *British Journal for the Philosophy of Science*, 67, 297–335.

Tal, E. (2017a). A Model-Based Epistemology of Measurement. In Mößner, N., and A. Nordmann, eds., *Reasoning in Measurement*. Routledge, pp. 233–253.

Tal, E. (2017b). Calibration: Modelling the Measurement Process. *Studies in History and Philosophy of Science*, 65–66, 33–45.

Worrall, J. (1989). Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories. In Gooding, D., T. Pinch, and S. Schaffer, eds., *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge University Press, pp. 135–157.

https://doi.org/10.1017/9781108885676 Published online by Cambridge University Press

Cambridge Elements

The Philosophy of Physics

James Owen Weatherall

University of California, Irvine

James Owen Weatherall is Professor of Logic and Philosophy of Science at the University of California, Irvine. He is the author, with Cailin O'Connor, of *The Misinformation Age: How False Beliefs Spread* (Yale, 2019), which was selected as a *New York Times* Editors' Choice and Recommended Reading by *Scientific American*. His previous books were *Void: The Strange Physics of Nothing* (Yale, 2016) and the *New York Times* bestseller *The Physics of Wall Street: A Brief History of Predicting the Unpredictable* (Houghton Mifflin Harcourt, 2013). He has published approximately fifty peer-reviewed research articles in journals in leading physics and philosophy of science journals and has delivered over 100 invited academic talks and public lectures.

About the Series

This Cambridge Elements series provides concise and structured introductions to all the central topics in the philosophy of physics. The Elements in the series are written by distinguished senior scholars and bright junior scholars with relevant expertise, producing balanced, comprehensive coverage of multiple perspectives in the philosophy of physics.

Cambridge Elements

The Philosophy of Physics

Elements in the Series

Global Spacetime Structure
JB Manchak

Foundations of Quantum Mechanics
Emily Adlam

Physics and Computation
Armond Duwell

Epistemology of Experimental Physics
Nora Mills Boyd

A full series listing is available at: www.cambridge.org/EPPH