

ARTICLE

Observations, Experiments, and Arguments for Epistemic Superiority in Scientific Methodology

Nora Mills Boyd¹ and Dana Matthiessen²

¹Philosophy Department, Siena College, Loudonville, NY, USA and ²Minnesota Center for Philosophy of Science, Minneapolis, MN, USA

Corresponding author: Nora Mills Boyd; Email: nboyd@siena.edu

(Received 12 May 2022; revised 28 May 2023; accepted 03 August 2023; first published online 10 August 2023)

Abstract

This article argues against general claims for the epistemic superiority of experiment over observation. It does so by dissociating the benefits traditionally attributed to experiment from physical manipulation. In place of manipulation, we argue that other features of research methods do confer epistemic advantages in comparison to methods in which they are diminished. These features better track the epistemic successes and failures of scientific research, crosscut the observation/experiment distinction, and nevertheless explain why manipulative experiments are successful when they are.

1. Introduction

Philosophers of science have inherited a distinction between *observation* and *experiment* that purports to track an epistemic difference.¹ The distinction turns on understanding experiment as active manipulation. In contrast, observation is cast as characteristically nonmanipulative. By virtue of this difference, some claim that experimentation is epistemically superior to observation, all things considered. This has two consequences: First, it entails that a researcher deciding between physically nonmanipulative or manipulative methods that are in other ways equal should opt for the manipulative. Second, it entails that sciences in which researchers lack the ability to physically manipulate their targets of inquiry are in a worse epistemic position than those who can. We will argue against these claims. Although there can be practical grounds for drawing a conventional distinction between observation and experiment, any such distinction does not, as a general matter, track a difference in the epistemic merits of scientific methods.

To better understand scientific methodology, we propose shifting the focus from physical manipulation, as highlighted by the observation/experiment distinction, to an alternate set of features that can crosscut this distinction. This accounts for the

Both authors contributed equally to the conception and writing of this paper. The author name order does not reflect intellectual priority.

¹ The contemporary distinction was not fully codified in scientific discourse until the turn of the 19th century (Daston 2011; Schickore 2019).

epistemic boon of manipulation where appropriate, but without attributing the success of these methods to manipulation per se. In that sense, our approach gets at more basic features of empirical methods to account for their superiority.

In section 2, we provide evidence that a view of experimentation as epistemically superior to observation recurs in the philosophy of science literature and identify common underlying assumptions. In section 3, we state an argument for this view in (logically) stronger and weaker forms and dismiss the stronger form. Section 4 argues against the claim that experiment is “in principle” superior to observation (including under some *ceteris paribus* assumption). Section 5 defends our alternate set of features.

2. The traditional view

John Herschel’s pioneering methodological treatise defined observation as “noticing facts as they occur, without any attempt to influence the frequency of their occurrence,” as opposed to “putting in action causes and agents over which we have control, and noticing what effects take place; this is EXPERIMENT” (Herschel 1831, 76). The core of this distinction—passive observation opposed to manipulative experiment—has been preserved to the present, although not without contestation.

Perović (2021) outlines two camps debating the observation/experiment distinction. Hacking (1983) claims these are separate kinds of activities: experiments are paradigmatically creations of phenomena—regularly occurring discernible effects—isolated from the causal complexity of the world; observations are roughly equivalent to detection by an instrument. Alternatively, Gooding (1992) and Malik (2017) argue that no general distinction can be drawn between these activities. Perović joins Brandon (1994) in placing these concepts on a continuum, where a more manipulative form of interaction with a system is the hallmark of experiment: “[t]he notion of experiment is certainly identified by the heightened substantial extent of manipulability in investigations” (Perović 2021, 9–10). We agree with Perović that a taxonomic distinction between observation and experiment, drawn with reference to the ability to physically manipulate a target, is likely to hold up for a wide range of cases and can serve nonepistemic aims. Our claim is that this distinction has no general *epistemological* significance for scientific practice.

Perović grants that “the high manipulability at one end of the continuum is not always epistemically superior to low-manipulability observations” (2021, 14) and indeed that experimenting “may be epistemically inferior or even detrimental to our knowledge of the desired phenomenon” in some contexts (15). For many authors, however, such claims are exceptions to a general rule holding experimentation superior to observation for confirmatory purposes. Currie and Levy (2019) identify this as the “traditional view,” wherein experiments are “seen as a privileged method of bringing the empirical and the theoretical into contact When experimentation is infeasible—in cosmology, geology, much of evolutionary biology and other areas—this is seen as a barrier to progress” (1066).

Several variants of the traditional view are expressed by authors in recent articles. A very strong version of this claim is found in Hacking (1983, 1989), who claims that “[n]atural (experimental) science is a matter not of saving phenomena but of creating phenomena [...] But in astrophysics we cannot create phenomena, we can only save them” (Hacking 1989, 578) and that indeed, “astronomy is not a natural science at all” (577).

Okasha (2011) has asserted the epistemic superiority of experiment on Bayesian grounds, arguing that only experimental manipulation allows one to confirm lawlike generalizations. From a causal modeling perspective, Zweir (2013) argues that the knowledge produced by observing a system and the knowledge produced by interventions on that system are “essentially different” (663), in part because observation leaves the nature of causal connections between correlated variables underdetermined.

It is not uncommon, then, for philosophers of science to claim experimentation provides a generic epistemic advantage over observation.² Currie and Levy (2019) are exemplars, for whom “the traditional view is right: experiments are indeed a privileged means of confirmation” (1067). We will consider their argument further because it shares some core assumptions with the general view we aim to critique.

For Currie and Levy, *control* is the key to understanding experimentation’s epistemic privilege over observation.³ Control consists of three features: (i) isolation of an object of study from its natural environment in a way that leaves focal properties undisturbed, (ii) manipulations in which researchers causally interact with the relevant properties of an object to change those (and only those) properties, and (iii) the ability to repeat the experiment many times. They summarize: “An object is subjected to control when isolated from its natural environment and intervened upon in a replicable way” (Currie and Levy 2019, 1071).

A few points are worth noting. First, Currie and Levy describe manipulation in terms of a particular form of causal interaction—a researcher’s physical intervention on the target system. Newton altering the positions of glass prisms in his light experiments is their example. In this sense of manipulation, it is impossible to experiment on remote systems such as galaxies.⁴ Second, their focus is restricted to successful hypothesis-testing experiments. The features of control are meant to define an ideal: experiments at their very best. When experimental control is optimally realized, they claim, it outperforms observation as a means of generating evidence.

3. Arguments for epistemic superiority

How are the features of control taken to ensure the superiority of experiment? “Performing controlled manipulations generates fine-grained, discriminating information” (Currie and Levy 2019, 1067). We adopt the following working definition of epistemic superiority, which reflects this intuition:

ES. An empirical method X is epistemically superior to Y, with respect to a system of interest S, if X produces results that reliably discriminate between more hypotheses in question about S than Y.⁵

² Mikhalevich (unpublished manuscript) also cites work promoting this claim in her own defense of nonexperimental methods.

³ They define experiment as controlled manipulation of a specimen. Both successful observations and experiments require a typical specimen, so control is the crucial difference.

⁴ Consider an influential methodology text: “Nonmanipulable events (e.g., the explosion of a supernova) or attributes (e.g., people’s ages, their raw genetic material, or their biological sex) cannot be causes in experiments because we cannot deliberately vary them” (Shadish et al. 2002, 7).

⁵ Our ES is similar to Roush (2018), who compares experiment to simulation. We recognize that hypothesis discrimination is neither the only epistemic aim in science nor the only way to evaluate methods.

A method reliably discriminates between hypotheses when it regularly generates results consistent with a proper subset of a range of hypotheses in question. One method may be superior to another by, for instance, discriminating hypotheses at a finer grain.⁶

Judgments of epistemic superiority carry normative weight for scientific decision making. If method X is epistemically superior to Y with respect to S, then a scientist seeking to pare down hypotheses about S ought to choose method X. If experimental methods are generally superior to observational methods, this provides a strong reason for a scientist deciding between more or less manipulative methods to prefer the former.

The epistemic superiority of physically manipulative experiments may be defended by way of one (or both) of the following theses:

Strong Control. Only experiments allow for fine-grained control over the production of data in isolation from confounding factors.

Strong Causation. Only experiments allow for causal inferences from data.

If, per **Strong Control** or **Strong Causation**, experiments can produce finer-grained results or better distinguish between different hypotheses about the causal relations between components of the system under investigation, then they can discriminate between more hypotheses than methods that lack these features. This would make experiment epistemically superior to observation on account of the special role of manipulation in affording control and yielding causal knowledge.⁷

Typically, however, the theses are weakened to the following *ceteris paribus* forms:

CP Control. Experiments allow for more fine-grained control over the production of data in isolation from confounding factors, all other things being equal.

CP Causation. Experiments allow for better causal inferences from data, all other things being equal.

The reasons for this weakening are straightforward. Cases are readily identified where being able to physically manipulate a system is not strictly necessary for fine-grained control over unconfounded data (as in early cosmic ray research [Galison 1982]) or for causal inference (which can exploit statistical dependencies within large observational datasets). It is true that researchers will have to contend with confounders, and these may bedevil certainty with respect to particular causal relationships. Yet the threat of confounders also applies to manipulative methods, such as when a manipulation is imprecise or inadequately understood. Thus, manipulation, in the sense of physical interaction, is neither necessary nor sufficient for producing the empirical results associated with either fine-grained control or causal inference. It follows that both **Strong Control** and **Strong Causation** are untrue and fail to justify the superiority of experiment to observation along the lines of ES.

⁶ Consider two methods for measuring a physical constant: one discriminating between positive/negative numbers and the other between large/small positive/negative numbers. ES recommends the latter.

⁷ In cases of applied/synthetic science in which the aim of the research is to create or engineer a system, it is trivially true that manipulation is necessary. Our argument concerns experiments *in general*.

4. CP arguments for the epistemic superiority of experiment

There are contexts in which the practical superiority of observational over experimental investigation is obvious. Such cases are most familiar in fields of research where the research object is easily disrupted. Dian Fossey studied unperturbed gorillas, not because it was impossible to intervene but because intervening would have altered the very phenomena she aimed to investigate. A proponent of experiment's superiority could object that such examples are unfair because they ignore the "in principle" advantages of physical manipulation to emphasize practical difficulties in realizing them in particular cases. It is more charitable to view arguments for experiment's epistemic superiority as appeals to theses like **CP Control** or **CP Causation**, which locate the advantage of experiment in the quality of the results it tends to produce. We will consider versions of each thesis.

4.1 Observations and confirmatory power

Currie and Levy are plausibly cast as proponents of **CP Control**. "[C]ombined with isolation and manipulation," they write, "repeatability allows an experimenter to conduct rich explorations of her object by finely varying the experimental circumstances while generating many significant data" (2019, 1071). And they conclude: "Control allows the generation of bountiful fine-grained, relevant, data and evidence" (Currie and Levy 2019, 1088). Each of the three features of control can be understood to play a role here. The data are bountiful, in part, because the experiment is repeatable, such that researchers can effectively generate data at will. They are fine-grained because the ability to manipulate a target allows for subtle variations in independent variables. They are relevant because other variables that may interfere with the properties under inquiry have been prevented from interfering through isolation and precisely designed manipulation.

For Currie and Levy, these features of control coincide with successful experiments, in which scientists have a high degree of manipulative access to a target of inquiry.⁸ When manipulation renders the experimental object less representative vis-à-vis a natural target of interest, and thus conflicts with the external validity of a method's results, then researchers may opt for more observational methods. But they do so at an epistemic cost, losing the kind of data that comes with control:

If, in order to conduct a controlled investigation, I must change my object such that it is not (or is less) relevantly representative of my target, then my capacity to confirm hypotheses is diminished. Vice-versa, when control is sacrificed for specimenhood, the confirmatory significance of the results is similarly weakened. (Currie and Levy 2019, 1075).

A method yielding the kind of data Currie and Levy associate with control would do better, by **ES**, than a method yielding similar data that are less abundant, less fine-grained, or less relevant to the hypotheses in question, all other things being

⁸ We bypass the fact that researchers often manipulate one system to learn about another. Currie and Levy distinguish between the object and target of research for this reason. This calls attention to mediating inferences from object to target but does not affect our argument.

equal. It is evident, for example, that finer-grained data would allow one to discriminate between more hypotheses than coarser-grained data (see footnote 7). We agree that superior methods afford data of this sort, but we question whether these should be identified with experiment qua physical manipulation of a target. This is not simply a disagreement about what counts as an “experiment” but what features of a data-gathering setup correspond to what epistemic goods.

Why do observational methods purportedly have worse confirmatory power, *ceteris paribus*? Currie and Levy first cite Okasha (2011), who provides a Bayesian analysis of the confirmation of scientific laws, conceived of as universal generalizations of the form $(x)(Fx \rightarrow Gx)$. Okasha claims that observation can only provide conjuncts of properties—*Fa* & *Ga*—as evidence, whereas manipulation via experiment allows one to produce an *Fa*, incorporate this into one’s background knowledge, and only then test to see whether it is also *Ga*. Unlike the conjunct properties, the latter sequence must increase the probability of the generalization.

Is this an adequate way of distinguishing observation from manipulative experiment? Imagine an observation that is informed by background knowledge in a manner that follows Okasha’s account of experimental procedure, one in which scientists know that *a*’s within some domain have property *F*, know the typical properties of *a*’s in other domains, and intentionally position themselves to see whether *a*’s with *F* have other unique properties. Okasha notes that this would confer similar confirmatory power as experiment. He invokes the category of “organized observation” in response to such a hypothetical involving an ornithologist, adding, “[r]ather than taking this to show that experimentation is not the only way of getting into the epistemic situation in question . . . I suggest that we instead conclude that the ornithologist did indeed perform an experiment, of a rudimentary sort” (Okasha 2011, 229).

Which contemporary observational sciences depend on naive non-“organized” observation rather than Okasha-style experiments? Every scientific form of data gathering that we are familiar with involves deliberate agency qua “way of putting oneself into the right epistemic situation.” If this is all that is required for a method to reap the advantages of experiment for Okasha, then his analysis alone does not warrant the stronger conclusion that experiment qua physical manipulation of a target is especially confirmatory. On the contrary, we take this to be support for the claim that the goods commonly associated with experiment can be obtained by other means.

We can further motivate this claim (with an eye toward **CP Causation** as well as **Control**) by contrasting manipulation as the physical, causal alteration of a system, with manipulation understood as a kind of dependency structure discoverable within data, given appropriate warrant from background knowledge. Woodward’s (2003) manipulability account of causal explanation exemplifies this approach. In this view, we can identify properly causal relationships in data insofar as they afford the right kind of modeling. This may or may not involve a scientific agent physically altering a target system:

[I]t is heuristically useful to think of an intervention as an idealized experimental manipulation carried out on some variable *X* for the purpose of ascertaining whether changes in *X* are causally related to changes in some other variable *Y*. [. . .] [A]ny process, whether or not it involves human activities, will qualify as an intervention as long as it has the right causal characteristics. (94)

Woodward's idealized manipulation involves the same kind of surgical intervention that Currie and Levy associate with causal control: it alters an independent variable of interest (and only that variable), and it "breaks the arrows" between the dependent variable of interest and any other upstream causes, shielding it from extraneous influence. The important difference is that under the right circumstances, causal modelers can identify interventional patterns of this sort from observational data.⁹ As long as data of sufficient detail can be paired with sufficient background knowledge, these data can be found to bear the same kind of structures as those that result from paradigmatic experiments and thus allow for correspondingly fine-grained discrimination between relevant hypotheses. Characterizing manipulation in terms of dependency relationships between variables renders this notion applicable to remote systems, such that we can gain knowledge of "interventions" and give causal explanations of "past events and of large-scale cosmological events" despite the fact that we cannot physically alter them (Woodward 2003, 10). Natural and social scientists can draw on background knowledge and data modeling to extract detailed interdependencies from observational datasets (cf. Morgan 2013; Bromham 2016). This contradicts the claim that experimental control gains generic confirmatory advantage over observation by generating data that bear interventional structure.

4.2 Frequency and CP claims are underspecified

Advocates of **CP Control** might respond to this in at least two ways. First, they might defend experimental superiority by claiming that it is *more often* the case that experimentation will yield high-quality results. Currie and Levy, for example, argue that so-called natural experiments rarely repeat, and because repetition "allows an experimentalist to reliably examine fine grained distinctions between variables—how one variable changes across a range of alterations to another variable" (2019, 1087); this means that observations of natural processes will rarely yield data allowing for comparatively fine-grained discrimination between hypotheses. This argument depends on exactly *which* natural processes fall within its scope. Some processes, like mass extinctions, are extremely rare; others, like some evolutionary patterns, recur on a regular basis; still others, like cosmic ray events, occur continuously, at a higher rate than humans could readily match through experimental production. We question how easily one can generalize over the vast range of phenomena beyond physical manipulation; identify those of scientific interest; and claim they rarely afford the kind of data that can be modeled in terms of, say, idealized manipulations. Without a more thorough accounting of such processes and our means of observing them, we find generalizations about their relative rarity unconvincing.

Second, advocates of **CP Control** or **Causation** might emphasize that they are talking about ideal experiments—scenarios that fully realize the features that make successful experiments so successful. For these purposes, it would be a mistake to compare experiments to observations in circumstances where the former are practically infeasible. Nor should we compare ideal experiments to sloppy observations. We should compare the best observations to the best experiments,

⁹ For an idealized manipulation to determine a causal relation between X and Y , one needs an observed variable W that either directly causes X or shares a common cause with X , is not an effect of X , and is not a direct cause of Y .

all else being equal. In *these* cases, the argument goes, experiments are epistemically superior.

This argument also requires further specification. What exactly is the “all else” that is equal here? We might assume a scenario in which a researcher is deciding between two methods of inquiry—one that involves physical manipulation and another that does not—where both concern the same target, where both are used to test the same set of hypotheses, and where the researcher has equally refined understanding of both the instruments and techniques involved in generating data from this target and how these interact with target properties of interest. We might go further and assume that each method provides data of an equally fine grain, that the data from each are equally relevant to the research questions, that each produces similar quantities of data, and that each affords a similar range of variation in target properties and measurable background conditions . . . If we held all features pertaining to data quality equal between the two methods, and held that they *only* differed on the matter of physical manipulation, it is not obvious to us that one method would discriminate between more hypotheses than another. Much depends on how equal we are trying to make which aspects of the two methods. For the *ceteris paribus* claim in **CP Control** or **Causation** to work, there must be some special differences in kind or degree with respect to the data that physical manipulations afford over the best observations, such that they can discriminate between more hypotheses. We now turn to several strong arguments that identify such features.

4.3 Randomization

Authors defending **CP Causation** in particular may argue that the capacity to schedule interventions allows experimenters to better manage potential confounds in data than observers. Per this argument, the epistemic superiority of experiment is exemplified by the advantages of controlled trials vis-à-vis observational studies that depend on correlational data alone. Considering an observational method and one carried out in a lab, we can suppose that the quality of data collection is equally reliable in each case, the understanding of the techniques equally sophisticated, the variations in treatment equally subtle, and so on, yet still think there is a distinct advantage to the lab-based approach. This advantage comes from the fact that researchers in the lab can manage the timing and target of treatments. This allows them to rule out relevant confounders that might be difficult to address in the wild and thereby improves their ability to secure causal hypotheses beyond the means available through observation (entailing superiority by **ES**). Some authors have gone as far as to identify this form of manipulative control as that “which makes a true experiment possible” (Campbell and Stanley 1963, 34).

The clearest argument for this form of superiority is found in the rationale for randomized control trials (RCTs). The basic idea is this:¹⁰ If researchers can schedule when and to whom some treatment occurs, then they can use a chance process to randomly assign subjects to different treatment groups. As long as this process provides the correct proportion of subjects from every relevant subgroup of the population, then factors other than the treatment that could affect individual outcomes, observed and

¹⁰ This explanation is largely drawn from Shadish et al. (2002).

unobserved, are randomly distributed over the different groups. The exclusion of unobserved confounders is crucial because it means randomization can reliably secure causal inferences without requiring thorough knowledge of a potentially overwhelming number of factors. Such a situation is not available for studies in which the scheduling of interventions is impossible, and researchers must deal with or rule out hypothetical confounders individually. One could argue that studies with random assignment thus can rule out more hypotheses and yield more reliable causal inferences than those that do not, and thus they are preferable in circumstances where researchers have a choice between distinct methods that are otherwise equal.

This argument makes a clear case for the value of randomization procedures, the implementation of which may depend on manipulative control. Yet the scope of its conclusion needs to be clarified. First, arguments favoring randomized experiment are strongest when applied to research contexts where the background knowledge of confounders is weakest. The need to control for the influence of confounders is more pressing the less one understands their number, variety, respective degree of influence, and susceptibility to measurement. Dealing with confounders is a vital concern across the sciences, but the situation is not uniform throughout. In many cases, potential confounders can be tamed by well-informed reasoning about the data-generating process.

The most important effects of random assignment are (i) elimination of researcher bias in the assignment process and, relatedly, (ii) assurance that the subgroups of the studied population are, on average, similar in all relevant respects other than the properties of interest.¹¹ In areas of research unafflicted by the kinds of confounders encountered in field experiments, RCTs are not needed to achieve these ends. Even in field experiments, there are alternative methods for the removal of selection bias (Worrall 2002). Moreover, the statistical advantages of having an unbiased sample are not lost on observational scientists, even though they do not perform RCTs. Raimann et al. (2005), for example, carefully selected a population of active (highly radio-luminous) galaxies and a “control” sample of nonactive galaxies. The control sample was carefully defined to illuminate differences between active and nonactive galaxies of similar morphological types and absolute magnitudes and avoid introducing selection effects that would bias the sample, for example, by using a particular spectral characteristic to select the control group (1241). This study illustrates how populations can be usefully defined in astronomy by drawing on background knowledge, that is, how researchers can obtain the epistemic benefits of randomization through other means.¹²

Arguments for the superiority of manipulative methods that enable random assignment are most convincing with respect to the scientific contexts that helped birth these methods, particularly human field experiments. Even then, the benefits of RCTs need to be qualified.¹³ Noting the advantages that RCTs can bring to field

¹¹ These underlie the only unflawed argument for randomization, according to Worrall (2002).

¹² See also Morgan (2013) on “as if” randomization. Climenhaga et al. (2021) argue that noise interference methods (NIMs) provide an alternative to RCTs for successful causal inference, including where RCTs would be physically impossible. They add, “[n]ot only is no intervention necessary for NIMs, nothing even like an intervention is necessary” (164).

¹³ Howson and Urbach (1989) and Deaton and Cartwright (2018a, 2018b) scrutinize the assumptions involved in determining that a given randomization procedure does in fact eliminate all confounders and the circumstances in which they may or may not be warranted. See, for example, Bluhm and Borgerson (2011) for further arguments about standards of evidence and the epistemology of RCTs.

research does not deliver the epistemic superiority of experiment over observation in general. More specifically, the ability to randomize treatments through physical manipulation does not render experiments superior to observation, all other things being equal, if the benefits of randomization can be obtained by other means available to observational methods. In scenarios where experimental interventions are possible, they only confer superiority when these other methods are assumed to be unavailable or uniformly worse in their results. Such assumptions require more detailed comparative argument than is provided in the literature.

4.4 Causal inference and counterfactuals

Even in idealized circumstances where both observation and carefully controlled physical interventions are possible, experiment is not by default superior to observation for purposes of causal discovery. Spirtes et al. (2000, ch. 9) survey a series of formal cases of causal inference in which both experimental and observational methods are capable of distinguishing between three hypotheses: a variable X (correlated with Y) causing Y , being caused by Y , or being correlated as a result of a common cause W . “Inferences to causal structure,” they write, “are often more informative when experimental data is available, not because causation is somehow logically tied to experimental manipulations, but because the experimental setup provides relevant causal knowledge that is not available about non-experimental data” (Spirtes et al. 2000, 260). However, these authors demonstrate that when the proper background knowledge or measurement procedures are available, causal knowledge of equal quality can be secured without a controlled experiment. For example, if there are variables U and V , which are already known to have the same causal relation to X and Y as interventions on these variables,¹⁴ then the relation between X and Y can be inferred from data. Going further, Spirtes et al. show that two causal structures, one where X causes Y and another where both have an unmeasured common cause W , can be distinguished observationally when embedded in a larger structure of measurable variables but cannot be distinguished by an experimental intervention on X .¹⁵ In other words, they present an ideal case where observation can distinguish between two hypotheses that experiment *cannot*. Here, ES would rule in favor of observation. They conclude that “the advantages of experimental procedures in identifying (as distinct from measuring) causal relations need to be recast” (Spirtes et al. 2000, 270).

Still, one might claim that some kinds of data are available *only* through physical manipulation, and this secures its superiority. Specifically, physical manipulation allows us to alter a system or its environment in ways that do not naturally occur and gather data from this. Therefore, experimentation allows us to gather more data than observation—in particular, data that are better for making counterfactual claims about a system of interest and thereby narrow down hypotheses about causal dependencies. This argument is another way of framing Currie and Levy’s remarks about the importance of being able to repeatedly manipulate through experiment:

¹⁴ That is, U can produce a known distribution on X in a manner that is independent of any other influence on X and where all causal paths between U and variables other than X must go through X .

¹⁵ This is because an intervention on X in formal causal modeling breaks the dependency between X and any other variable. Because X is also dependent on W in this case, the intervention destroys important information for determining the true structure.

“Repetition underwrites both inferences from data to phenomena and establishing a result’s external validity vis-à-vis a target. Generating such knowledge often requires an enormous number of variations, test runs, and so forth. Although natural experiments can play a confirmatory role, their power is limited by a lack of finely varied repetitions” (2019, 1087).

For some scientific pursuits, this strikes us as a clear-cut argument in favor of experiment.¹⁶ But the general claim that the extra data produced through interventions are *better* simply because they expose a system to a broader range of conditions needs to be considered in light of the kinds of hypotheses that are “in question” for researchers, as we put it in *ES*. Even when exposing a system to artificial conditions, scientists are routinely interested in hypotheses concerning its behavior beyond the lab. For this purpose, more experimental data is not always better—many laboratory contrivances may alter target behavior in a way that ruins the external validity of results, as in animal ethology. Similarly, when the hypotheses in question in a field have a high degree of historical specificity, as in molecular phylogeny, experiment is viewed as secondary to observational methods because these results only provide support for a general model of a process rather than the actual process that took place (O’Malley 2016). The ability to physically manipulate aspects of the target system and its environment does not provide an epistemic edge unless these alterations yield data that are well suited to the hypotheses in question within a given research program. Conversely, the inability to physically manipulate a target only entails worse confirmatory power than otherwise if observational data collection does not afford the right kind of modeling with respect to the hypotheses in question or is generally worse for investigating their details. The previous discussion of Spirtes et al. (2000) shows that this is not true. It is a matter that must be dealt with on a case-by-case basis rather than a generic advantage of manipulative methods. Finally, the assumption that manipulative methods yield data covering a broader range of conditions requires further qualification. There are certain research contexts in which nature’s variations outstrip the generative capacities of human artifice. Hence, it is remarked that the universe is the “poor man’s accelerator” because the universe naturally produces conditions that cannot be achieved in terrestrial experiments as a result of the sheer energy required.

But perhaps this misses the “in principle” nature of the argument. One could argue that if it *were* possible to perform experiments that are in effect impracticable, then scientists would have access to more, better data. To borrow the case that Jacquart (2020) discusses, if astrophysicists could physically create head-on collisions between actual compact galaxies and disk galaxies under various conditions, they might learn more about the formation and evolution of ring galaxies than they can without physically smashing galaxies into one another. Performing physically impossible experiments would deliver more results than mere observation, thereby making experiment epistemically superior to observation, in principle.

Our response is to question the value of this line of reasoning. If the epistemic superiority of experiment is *only ever* a matter of principle—that is, if there are no practical circumstances under which experimental method X can ever be carried out instead of observational method Y—then this alleged superiority loses its normative weight for scientific decision making. For this reason, we think discussions of epistemic

¹⁶ See footnote 8.

superiority as it bears on scientific practice should satisfy a *minimal pragmatic condition*: if X is epistemically superior to Y, then there must be some possible practical circumstances under which a researcher could choose X over Y.¹⁷ Perhaps the “in principle” argument can be made convincingly from a “God’s eye view,” but even if it can, it would not tell us anything useful for scientific decision making in practice. In particular, it would not imply that actually existing, or even nomically possible, observational sciences are generally worse off in their pursuit of highly informative data.¹⁸

We have considered several arguments that experiment is epistemically superior to observation because it allows for a degree of control that yields higher-quality results for hypothesis discrimination, all other things being equal. Some arguments for this claim, such as Okasha’s, characterize observation and experiment in terms that do not track a distinction of scientific interest. In fact, there are means for observing natural phenomena that, when supplemented with appropriate background knowledge, allow scientists to investigate the same kinds of intervention-based hypotheses as methods involving physical manipulation. We are skeptical of pro-experiment arguments that push against this point by an underspecified appeal to the rarity of such observational opportunities. Similarly, we question what is held “equal” in comparisons between ideal experiments and ideal observations. It is not obvious that a truly ideal observation is a priori worse off; formal results suggest that, for some hypotheses, observational methods can be superior. We acknowledge the advantages of RCTs in certain research contexts but argue that the methodological benefits of randomization and unbiased sampling are appreciated and utilized in observational sciences and, more importantly, that the parochial usefulness of RCTs does not amount to support of a generic claim for experiment’s epistemic superiority over observation. In most examples we can countenance, the superiority of a method depends on context-sensitive details of the kind of research being carried out. This includes cases where experiment allows researchers to contrive novel forms of data. This ability, when available, is only better if conducive to the hypotheses in question. When unavailable, the arguments in experiment’s favor are confined to “in principle” speculations with no normative force for science in practice. We leave it to the defenders of the epistemic superiority of experiment to offer further suggestions for the justification of that position. In contrast, our analysis thus far supports the conclusion that the distinction between observation and experiment does not track any generic epistemic difference.

5. What does make empirical methods epistemically superior

There are certain features of methods used for empirical data-gathering practices that do generally make an epistemic difference, even though this is not true of physical manipulation. By “methods,” we have in mind the suite of instruments and techniques that researchers use to produce, record, and process empirical data. Methods might differ in the manner in which they engage with a target, the physical instruments and forms of data analysis in use, and their associated background

¹⁷ Because the decisions we are concerned with are those being made by scientists seeking to understand our actual world, the relevant possibilities are nomic, or perhaps merely empirical (Norton 2022).

¹⁸ See section 2: it is the practical normative output of judgments of superiority that warrants our interest here.

knowledge. These differences individuate methods investigating the same target. For example, some experiments designed to measure the neutron lifetime pass a beam of neutrons through a region of known finite volume and detect neutrons and protons that exit the volume (Wietfeldt 2018, 7). In contrast, experiments using a version of the “bottle” method trap ultra-cold neutrons in a storage volume and measure how many have survived after a specified time has elapsed (Wietfeldt 2018, 13). Although these aim to measure the same quantity, each method introduces characteristic benefits and challenges.

Scientists often face choices about which method to choose to conduct empirical research. Methods with features such as higher signal clarity, better characterization of backgrounds, and/or increased discrimination and variability of precipitating conditions will be epistemically superior to alternatives in which these features are lower, worse, and/or diminished, all other things being equal. These are three salient epistemic features, or “parameters,” of empirical methods that do make an epistemic difference in terms of ES. These are likely not the only such features, but they are familiar and illustrative. In this section, we introduce these parameters with the aim of demonstrating, first, that they are relevant to claims comparing the general epistemic merits of empirical methods and, second, that these parameters crosscut the traditional distinction between observation and experiment. We argue that tracking the observation/experiment distinction is a worse way to judge the epistemic superiority of alternative empirical methods than an approach that directly concerns parameters such as signal clarity, characterization of backgrounds, and discrimination and variability of precipitating conditions. This argument has significant payoffs for philosophy of science, as we will discuss. We anticipate that a shift in focus from the observation/experiment distinction to contextual parameters of empirical research, such as the three we highlight here, will make the epistemology of empirical science more accurate, insightful, and applicable to science in practice.

5.1 Signal clarity

Whether a method is apt for investigating a particular system depends on how clear a signal researchers can expect to extract from this system using the method in question. The prospects of achieving a clear enough signal will depend on the extent to which the data-gathering setup is capable of recording the targeted behavior or properties of a system without interference from other contextual factors. When judging a method with respect to expected signal clarity, researchers may ask questions like these: “How sensitive is my apparatus to this property?” and “Are there regimes or conditions under which I can investigate this property where noise is sufficiently minimized?” A simple example is that longer exposure on a telescope will increase the signal with respect to certain kinds of noise. Similarly, cooling a detector generally reduces noise due to thermal fluctuations in the electronics, thereby lowering the noise floor and allowing for better signal clarity. For a more sophisticated example, consider the strategy that cosmologists use to study structure formation in the early universe by tracing the weak signal from neutral hydrogen using radio telescopes. Although informative, this weak cosmic signal competes with extremely bright foreground emission from our own galaxy. To increase their signal clarity, the cosmologists focus their investigation on a limited area of Fourier space in

which the foregrounds are relatively quiet (Liu and Shaw 2020, sec. 12.1.5). As a general rule, those methods will be preferred that employ instruments that are more responsive to the signal source and have physical setups that better screen or reduce the noise. In many cases, researchers quantify this feature of an empirical investigation in the signal-to-noise ratio (SNR).¹⁹

As an example of the general epistemic significance of signal clarity, consider trade-offs in neutrino research. Physicists can study neutrinos produced as by-products from nuclear power reactors, from the beta decay of tritium in a mass spectrometer, from highly enriched germanium crystals, from neutrinos produced in the sun and in supernovae, and so on. Each of these approaches can be realized in many different ways. Which approach is epistemically superior will depend in large part on what specific signal is sought and how strong that signal is expected to be, given the research context. For instance, many more neutrinos may be produced as reactor by-products than will arrive at terrestrial detectors from distant supernovae, but reactors also produce a lot of noise.

5.2 Characterization of backgrounds

Research that better characterizes background factors is generally epistemically superior. It is rare for a data-collection method to exclusively pick up the signal of interest. Data recorded from a system-instrument interaction typically include contributions from diverse causal factors. These “background” elements in data include contributions from the composition and/or operation of the apparatus, from other sources in the target’s environment, and/or from aspects of the target that are not of interest. Although backgrounds and noise both contribute unwanted elements to the data, they are functionally distinguishable in that backgrounds can be attributed more specifically to certain sources.²⁰

Bogen and Woodward’s (1988) distinction between data and phenomena is instructive here. For them, data are idiosyncratic, and their individual particular values do not call for theoretical explanations (305–6). We suggest this is so when statistically random variations are contributing to the values measured, but the researchers have no interest in tracking down the source of those variations—this is “noise” in the pejorative sense. However, scientists are often interested in isolating, explaining, and somehow dealing with unwanted contributors to the data. In his experiments on gravitational attraction, Henry Cavendish worried about contributions from air currents, magnetic forces, and distortions in his apparatus *in particular*, and he took concrete, ingenious, and intentional steps to rule out contributions from these various backgrounds (Galison 1987, 3). We might say, using Bogen and Woodward’s terminology, that particular backgrounds can temporarily become a phenomenon of interest to researchers on their way to investigating some other phenomenon or, alternatively, that familiarity with one phenomenon can be put to use in the investigation of another one in that the first can be recognized as an unwanted contributor to data collected in service of studying the latter. Thus, the

¹⁹ We take the distinction between signal and noise to be a functional and contextual one. See Bogen (2010) for a case where electroencephalogram (EEG) “noise” is treated as a research target rather than something to be removed.

²⁰ Craver and Dan-Cohen (forthcoming, sec. 5) make a similar distinction between confounds and noise.

common mantra of experimental physics: “Yesterday’s sensation is today’s calibration and tomorrow’s background.”

To remove these sorts of contributions and thus better isolate the signal within recorded data, researchers require means for identifying and canceling or subtracting irrelevant features from data. In some cases, this can be accomplished via physical shielding or other modifications to the apparatus or its environment. In others, researchers impose data cuts and/or masks or subtract background contributions from the data collected. In order to subtract backgrounds, researchers may attempt to measure them separately from the signal of interest and/or estimate them via modeling or simulation. As a general rule, those methods will be preferred that are accompanied by better means for measuring or calculating these background contributions to recorded data.

For further illustration, consider neutrino research again. The IceCube Neutrino Observatory deployed a massive array of photodetectors under a solid cubic kilometer of ice at the South Pole. The ice is the detector in this case: high-energy neutrinos from space interact with water molecules, and the light from that interaction is captured by the photodetectors. The success of this approach hinges on the purity of the ice; impurities introduce uncertainties in reconstructing the interaction. A next-generation detector has recently been proposed (called P-ONE) that would hang photodetectors in a larger patch of the Pacific Ocean, thereby creating a detector with a larger volume, which would be more likely to interact with high-energy neutrinos. However, the purity of the Pacific Ocean is a challenge for this approach—the researchers will have to figure out, for example, how to account for the (unwanted) contributions from bioluminescence (Resconi and P-ONE Collaboration 2021). Whether or not moving to a larger detector in the Pacific is actually epistemically superior to a smaller-scale experiment with higher-purity detector material will depend crucially on physicists’ success in characterizing and accounting for the emission of light by ocean organisms.

5.3 Discrimination and variability of precipitating conditions

A research method is generally epistemically superior insofar as it better discriminates and tracks the variability of precipitating conditions. The properties and behaviors of a system are produced (and can be modified) by certain precipitating conditions.²¹ To understand various properties and behaviors of a system, it is therefore beneficial to be able to distinguish between the different conditions that may affect them and track how these properties covary with such conditions. Astrobiologists wanting to know about the conditions under which life can form in the universe will derive greater epistemic benefit from research that includes conditions in exotic extraterrestrial environments as opposed to those that exclusively focus on planet Earth. These conditions are not “backgrounds” in the sense discussed earlier. Backgrounds are unwanted signals that contribute to the data collected, obfuscating or mimicking the signal of interest. In contrast, precipitating conditions are the conditions that produce the signal in the first place. As a general

²¹ Craver and Darden (2013) explicate precipitating conditions as “all of the many sets of conditions sufficient to make the phenomenon come about” (56).

rule, those methods will be preferred that allow for the discrimination of a larger number of precipitating conditions that vary over a wide range.

For example, the very high-energy neutrinos that IceCube detected could originate from a variety of astrophysical sources, including active galactic nuclei, supernovae, hypernovae, white dwarf mergers, and others (Mészáros 2017). An important part of the motivation for building larger next-generation detectors like P-ONE is to get high enough angular resolution to attribute the neutrinos detected to localized astrophysical sources, which researchers can also study using “multimessenger” approaches: investigating the same sources using optical, radio, gamma-ray, and gravitational wave astronomy (Halzen 2021). Higher resolution and multimessenger follow-up will allow researchers to learn about the variety of conditions that generate these high-energy neutrinos and study how differences in astrophysical source conditions affect the associated neutrino flux.

5.4 These context-specific parameters promote epistemic superiority

In general, higher signal clarity, better characterization of backgrounds, and higher discrimination and variability of precipitating conditions improve the epistemic outcomes of empirical research. Methods that better promote signal clarity increase the precision, accuracy, and confidence of an empirical result. Methods that better account for backgrounds prior to or after the recording of data will reduce systematic error. This also increases the accuracy of results by eliminating directional bias that shifts measurements away from the phenomenon or the postulated “true value” of the quantity of interest. Results that are more precise and accurate are better able to reliably discriminate between hypotheses because they can discriminate between finer ranges of values and can account for contributions of a wider range of confounding factors. By tracking how the variation of precipitating conditions correlates with the variations in a recorded signal, researchers can infer more complex relationships between the system of interest and its environment, allowing them to discriminate between more hypotheses about this system than they otherwise could. A method that can be used over a wider range of conditions may uncover new informative relationships, thereby reducing the blind spots of a more restrictive method. Those informative relationships can be used to better adjudicate between alternative hypotheses. In short, methods that do better according to these features are epistemically superior according to ES.

There is likely some overlap between the three features we have chosen to highlight. As we mentioned earlier, the distinction between noise that degrades signal clarity and a background that makes an unwanted contribution to data is largely a matter of our epistemic vantage point. If the source of the contribution has been, or can be, determined and dealt with, researchers treat it as a background. If the contribution is random and the source unknown, it is treated as mere statistical variation to be (hopefully) swamped by collecting more data and improving the SNR. Variations in precipitating conditions may contribute different noise levels or backgrounds. Although these features are interconnected, we nevertheless suggest that they are worth characterizing separately because they are often considered separately in scientific decision-making contexts. The error budget for empirical research will often be broken down into “statistical” and “systematic” components.

Researchers can wonder whether it is possible to investigate their subject matter under new precipitating conditions and then wonder what sort of backgrounds they might have to contend with in those cases.

5.5 Relation to the observation/experiment distinction

Note that each feature we've highlighted does not covary with the common distinction between observation and experiment. Researchers can refine the sensitivity of a detector, measure and remove background contributions to data, or study a phenomenon under a wider range of precipitating conditions without thereby making research more experimental, in the sense of increasing their manipulative access to a target. Likewise, opting for a less manipulative method need not entail a relative privation of these features. As we saw earlier, even though neutrinos can be sourced from terrestrial reactors under human control, it can be advantageous to study space-born neutrinos for a variety of reasons, including noise reduction. Some observers instead worry about how experiment may *worsen* results, distorting target signals and multiplying confounds through manipulative interventions. In short, our parameters crosscut the observation/experiment divide.

Are such cases exceptions to a general rule, according to which an increase in manipulative access *usually* or *tends* to yield better signal clarity, characterizations of background, or discrimination and variability in precipitating conditions? The features that Currie and Levy associate with controlled manipulations, for instance, appear to have a direct relation to these parameters. Isolation allows for the reduction of background contributions; targeted interventions may do the same while increasing signal clarity; repeatable manipulations can be done under varying conditions. Again, we agree that these procedures yield high-quality results; they are an important part of what makes the best experiments so successful. But there are parallel procedures in the best observational methods (e.g., rich data modeling paired with background knowledge) that achieve the same ends by different means, which make these methods equally successful in their domains. The success of some experiments is indeed due to manipulative control of a target, and observational methods may lack this feature, but this does not mean successful observations are worse off than successful experiments. And again, further arguments are required to judge the relative frequency of conditions under which the best experiments obtain in comparison to the best observations. As it stands, we claim physical manipulation is a red herring.

Our features allow us to explain the appeal of experiments when they do work, without attributing this to physical manipulation per se. In cases where researchers do actually face a choice between a more observational research method and a more experimental one, and in which the experimental option does happen to be epistemically superior, the reason for this is often because it increases one or more of the previously described parameters in that context. That is, the (local) epistemic superiority of experiment, when it is indeed superior, is derivative of the power of the features that we have highlighted. Although upping the manipulation of a system via experiment does not generally induce epistemic superiority in practice, upping features such as signal clarity, characterization of backgrounds, and the discrimination and variability of precipitating conditions generally does, and in some cases, this can be accomplished by increased manipulation of the target. For example, in cases

where medical researchers prefer controlled trials to intervention-free population studies on epistemic grounds, we claim that this preference is explained by the fact that the trials (say) allow for better elimination of confounds and discrimination of precipitating conditions than the studies would. Noting these epistemic benefits need not commit us to anticipating that experimental meddling will always or even usually improve one's epistemic lot. However, if we attribute the epistemic advantages to the experimental character of the research *per se*, then we risk committing ourselves to that mistaken inference.

5.6 Philosophical payoffs

The approach for which we advocate has further philosophical payoffs. One is that by refocusing on the epistemic power of more fine-grained variants in research methods, such as signal clarity, rather than the gross categories of “observation” and “experiment,” we avoid mistakenly dismissing the value of whole fields of scientific research on account of their “observational” nature. Ian Hacking notoriously (and wrongfully) disparaged the scientific character of the entire fields of astronomy and astrophysics as mere “saving the phenomena” (1989, 557–58). This sort of judgment is not available once we dismiss the observation/experiment distinction as a red herring for the epistemology of empirical research. Rather than lamenting the fact that a field is characteristically observational, our approach presses philosophers of science to investigate what clever approaches scientists have actually leveraged in practice in that field to make epistemic progress and what challenges—and hopes for their resolution—remain.

Another payoff of our approach is that it helpfully draws attention to where the epistemic “action” is in philosophical case studies. We contend that asking the question, “Is this research an observation or an experiment?” is not generally going to be particularly illuminating. In contrast, tracking features of the sort we have countenanced will be generally informative of the epistemic pitfalls and successes of empirical research. To take just one example, Boyd (2023) has argued that philosophical investigation of the methodology and epistemology of laboratory astrophysics is hindered by predicating that investigation on the observation/experiment distinction. Boyd's central case study is an example of laboratory astrophysics—astrophysics research conducted in a terrestrial laboratory. However, what makes or breaks the epistemology of this particular case does not depend on its character *vis-à-vis* observation versus experiment. As it happens, the research is experimental, but noting as much is not useful for understanding the opportunities it affords and the challenges it faces. Instead, Boyd argues that it is once we conduct a more fine-grained analysis of the data-generating process in this case that we see a problem for the interpretation of the results of the experiment provided by the researchers themselves. Thus, the more fine-grained approach can help philosophers of science more accurately direct their normative contributions to science in practice.

Attending to fine-grained details, like our parameters described earlier, also offers a way to deepen our analyses of epistemological terms of art in philosophy of science. One relevant example is “background knowledge,” which has been variously employed across the literature on inductive inference (e.g., Alexander 1958; Popper 1963; Longino 1990; Okasha 2001) and has even been called “the means” by which new knowledge of nature is acquired (Shapere 1982, 516), albeit with little explication or

agreement on what exactly is meant by this term. Employing umbrella concepts such as “background knowledge” (or its cousin “auxiliary hypotheses”) without further analysis homogenizes the diverse resources scientists draw on in producing empirical results. As long as these are reduced to a one-dimensional “background” status, the texture of scientists’ practical reasoning and activities will remain undertheorized. We submit that further articulation and discussion of the contextual parameters that make an epistemic difference within scientific methodology can bring out the heterogeneity underlying “background knowledge” in empirical research; highlight and specify the distinct contributions this knowledge makes to scientific inquiry; and challenge ideas of this knowledge as unproblematic, passively assumed, or otherwise playing a back-seat role in the course of investigation.

6. Concluding remarks

Our list of three parameters is not meant to be exhaustive; we chose these ones because they strike us as uncontroversial alternatives to thinking along the observation–experiment axis. Identifying the substantive roles of such features is helpful for appraising the epistemic significance of existing scientific research but also for informing decisions about what research ought to be conducted next and for choosing between available methods.

In this article, we have added arguments to the extant critiques of an epistemically significant distinction between experiment and observation. We argued that physical manipulation is not necessary for achieving fine-grained control and generating causal knowledge. We also articulated a pair of subtler “in principle” arguments for the epistemic superiority of experiment. However, we argued that even these more nuanced versions fail to support a generic claim to epistemic superiority relevant to decisions that scientists make in practice. Finally, we have identified some features of empirical data-gathering practices that we argue do generally confer epistemic superiority, that crosscut the traditional distinction between observation and experiment, and that can help to explain why manipulative experiments are successful when they are. Philosophers of science should shift their focus away from the distinction between observation and experiment toward more fine-grained features that are more informative for the epistemology of empirical research.

Acknowledgements. We are especially grateful for feedback from Kareem Khalifa, Adrian Currie, and the anonymous referees.

References

- Alexander, Henry Gavin. 1958. “The Paradoxes of Confirmation.” *British Journal for the Philosophy of Science* 9 (35):227–33. <https://doi.org/10.1093/bjps/IX.35.227>
- Bluhm, Robyn, and Kirstin Borgerson. 2011. “Evidence-Based Medicine.” In *The Handbook of Philosophy of Science. Volume 16: Philosophy of Medicine*, edited by Fred Gifford, 203–38. Amsterdam: Elsevier.
- Bogen, James, and James Woodward. 1988. “Saving the Phenomena.” *Philosophical Review* 97 (3):303–52. <https://doi.org/10.2307/2185445>
- Bogen, Jim. 2010. “Noise in the World.” *Philosophy of Science* 77 (5):778–91. <https://doi.org/10.1086/656006>
- Boyd, Nora Mills. 2023. “Laboratory Astrophysics: Lessons for Epistemology of Astrophysics.” In *Philosophy of Astrophysics: Stars, Simulations, and the Struggle to Determine What Is Out There*, edited by Nora Mills Boyd, Siska de Baerdemaeker, Keven Heng, and Vera Matarese, 13–32. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-26618-8_2

- Brandon, Robert. 1994. "Theory and Experiment in Evolutionary Biology." *Synthese* 99 (1):59–73. <https://doi.org/10.1007/BF01064530>
- Bromham, Lindell. 2016. "Testing Hypotheses in Macroevolution." *Studies in History and Philosophy of Science* 55:47–59. <https://doi.org/10.1016/j.shpsa.2015.08.013>
- Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.
- Climenhaga, Nevin, Lane DesAutels, and Grant Ramsey. 2021. "Causal Inference from Noise." *Noûs* 55 (1):152–70. <https://doi.org/10.1111/nous.12300>
- Craver, Carl F., and Talia Dan-Cohen. Forthcoming. "Experimental Artefacts." *British Journal for Philosophy of Science*.
- Craver, Carl F., and Lindley Darden. 2013. *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago: University of Chicago Press.
- Currie, Adrian, and Arnon Levy. 2019. "Why Experiments Matter." *Inquiry* 62 (9–10):1066–90. <https://doi.org/10.1080/0020174X.2018.1533883>
- Daston, Lorraine. 2011. "The Empire of Observation, 1600–1800." In *Histories of Scientific Observation*, edited by Lorraine Daston and Elizabeth Lunbeck, 81–114. Chicago: University of Chicago Press.
- Deaton, Angus, and Nancy Cartwright. 2018a. "Reflections on Randomized Control Trials." *Social Science & Medicine* 210:86–90. <https://www.sciencedirect.com/science/article/pii/S0277953618302132?via%3Dihub>
- Deaton, Angus, and Nancy Cartwright. 2018b. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210:2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Galison, Peter. 1982. "The Discovery of the Muon and the Failed Revolution against Quantum Electrodynamics." *Centaurus* 26 (3):262–316. <https://doi.org/10.1111/j.1600-0498.1982.tb00666.x>
- Galison, Peter. 1987. *How Experiments End*. Chicago: University of Chicago Press.
- Gooding, David. 1992. "Putting Agency Back into Experiment." In *Science as Practice and Culture*, edited by Andrew Pickering, 65–112. Chicago: University of Chicago Press.
- Hacking, Ian. 1983. *Representing and Intervening*. New York: Cambridge University Press.
- Hacking, Ian. 1989. "Extragalactic Reality: The Case of Gravitational Lensing." *Philosophy of Science* 56 (4):555–81. <https://doi.org/10.1086/289514>
- Halzen, Francis. 2021. "High-Energy Neutrinos from the Cosmos." *Annalen der Physik* 533 (11):2100309. <https://doi.org/10.1002/andp.202100309>
- Herschel, John Frederick William. 1831. *A Preliminary Discourse on the Study of Natural Philosophy*. London: Longman, Rees, Orme, Brown, Green, and John Taylor.
- Howson, Colin, and Peter Urbach. 1989. *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.
- Jacquart, Melissa. 2020. "Observations, Simulations, and Reasoning in Astrophysics." *Philosophy of Science* 87 (5):1209–20. <https://doi.org/10.1086/710544>
- Liu, Adrian, and J. Richard Shaw. 2020. "Data Analysis for Precision 21cm Cosmology." *Publications of the Astronomical Society of the Pacific* 132 (1012):062001. <https://doi.org/10.1088/1538-3873/ab5bfd>
- Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.
- Malik, Saira. 2017. "Observation versus Experiment: An Adequate Framework for Analysing Scientific Experimentation?" *Journal for General Philosophy of Science* 48 (1):71–95. <https://doi.org/10.1007/s10838-016-9335-y>
- Mészáros, Peter. 2017. "Astrophysical Sources of High-Energy Neutrinos in the IceCube Era." *Annual Review of Nuclear and Particle Science* 67:45–67. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Mikhalevich, Irina. "Intervention and Experiment." Unpublished manuscript.
- Morgan, Mary S. 2013. "Nature's Experiments and Natural Experiments in the Social Sciences." *Philosophy of the Social Sciences* 43 (3):341–57. <https://doi.org/10.1177/0048393113489100>
- Norton, John. 2022. "How to Make Possibility Safe for Empiricists" In *Rethinking the Concept of Law of Nature: Natural Order in the Light of Contemporary Science*, edited by Yemina Ben-Menahem, 129–59. Cham, Switzerland: Springer. <https://doi.org/10.1007/978-3-030-96775-8>
- Okasha, Samir. 2001. "What Did Hume Really Show about Induction?" *Philosophical Quarterly* 51 (204):307–27. <https://doi.org/10.1111/1467-9213.00231>
- Okasha, Samir. 2011. "Experiment, Observation and the Confirmation of Laws." *Analysis* 71 (2):222–32. <https://doi.org/10.1093/analys/anr014>

- O'Malley, Maureen. 2016. "Histories of Molecules: Reconciling the Past." *Studies in History and Philosophy of Science Part A* 55:69–83. <https://doi.org/10.1016/j.shpsa.2015.09.002>
- Perović, Slobodan. 2021. "Observation, Experiment, and Scientific Practice." *International Studies in the Philosophy of Science* 34 (1):1–20. <https://doi.org/10.1080/02698595.2021.1978038>
- Popper, Karl. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge and Kegan Paul.
- Raimann, Daniel, Thaisa Storchi-Bergmann, Hernan Quintana, Richard Hunstead, and Lutz Wisotzki. 2005. "Stellar Populations in a Complete Sample of Local Radio Galaxies." *Monthly Notices of the Royal Astronomical Society* 364 (4):1239–52. <https://doi.org/10.1111/j.1365-2966.2005.09665.x>
- Resconi, Elisa, and P-ONE Collaboration. 2021. "The Pacific Ocean Neutrino Experiment." In *Proceedings of 37th International Cosmic Ray Conference—PoS (ICRC 2021)*, vol. 395. <https://pos.sissa.it/395/024/pdf>
- Roush, Sherrilyn. 2018. "The Epistemic Superiority of Experiment to Simulation." *Synthese* 195 (11):4883–906. <https://doi.org/10.1007/s11229-017-1431-y>
- Schickore, Jutta. 2019. "The Place and Significance of Comparative Trials in German Agricultural Writings around 1800." *Annals of Science* 78 (4):484–503. <https://doi.org/10.1080/00033790.2021.1958008>
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Causal Inference*. Boston: Houghton Mifflin.
- Shapere, Dudley. 1982. "The Concept of Observation in Science and Philosophy." *Philosophy of Science* 49 (4):485–525. <https://doi.org/10.1086/289075>
- Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.
- Wietfeldt, Fred E. 2018. "Measurements of the Neutron Lifetime." *Atoms* 6 (4):70. <https://doi.org/10.3390/atoms6040070>
- Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.
- Worrall, John. 2002. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69 (S3):S316–30. <https://doi.org/10.1086/341855>
- Zwier, Karen R. 2013. "An Epistemology of Causal Inference from Experiment." *Philosophy of Science* 80 (5):660–71. <https://doi.org/10.1086/673901>

Cite this article: Boyd, Nora Mills and Dana Matthiessen. 2024. "Observations, Experiments, and Arguments for Epistemic Superiority in Scientific Methodology." *Philosophy of Science* 91 (1):111–131. <https://doi.org/10.1017/psa.2023.101>