# Bringing Thought Experiments Back into the Philosophy of Science

**Adrian Currie and Arnon Levy**

To appear in *Studies in the History and Philosophy of Science*

## Abstract

To a large extent, the evidential base of claims in the philosophy of science has switched from thought experiments to case studies. We argue that abandoning thought experiments was a wrong turn, since they can effectively complement case studies. We make our argument via an analogy with the relationship between experiments and observations within science. Just as experiments and 'natural' observations can together evidence claims in science, each mitigating the downsides of the other, so too can thought experiments and case studies be mutually supporting. After presenting the main argument, we look at potential concerns about thought experiments, suggesting that a judiciously applied mixed-methods approach can overcome them.

### *1. Introduction*

This paper discusses the relationship between case studies and thought experiments in the philosophy of science. Briefly put, our thesis is that the relative neglect of thought experiments in recent philosophy of science is unfortunate, as they can mitigate some of the shortcomings of case studies. In particular, while case studies are an important avenue for making contact with scientific practice, they are also subject to (familiar, but serious) concerns stemming from cherry picking, interpretational biases, and the complications of handling real-life science. In contrast, thought experiments operate at a certain remove from practice, impugning their value in some respects. But they afford opportunities for simplification, isolation and manipulation, which work with case studies lacks. So we recommend the use of thought experiments and case studies in tandem. In short, we defend a *mixed-methods* approach to the philosophy of science[1].

[1] In the social sciences 'mixed-methods' applies to combinations of qualitative and quantitative approaches, so our usage extends it somewhat.

Our argument is made against the backdrop of the last few decades' shift toward practice-oriented work (e.g., Ankeny et al 2011). Whereas in the early and middle parts of the 20th century philosophers tended to trade in toy examples and thought experiments, recent work probes actual scientific practice to a much greater extent. This shift is coupled with a (welcome) broadening of philosophical horizons: from a focus primarily on formal aspects of theorizing, especially in fundamental physics, to a rich array of scientific targets, objects and processes. The appeal to case studies is perhaps the most visible symptom of the practice-turn.[2]

We understand thought experiments to involve simplified, abstract description of scenarios designed to generate a focused judgement concerning circumstances relevant to normative questions in the philosophy of science. In contrast, we take case studies to be fairly detailed descriptions of particular episodes of scientific practice, drawn either from contemporary science or its history. These days, case studies are perhaps the main source of evidence for philosophers of science, a dramatic methodological shift from the mid-20th Century. Consider scientific explanation, an example we will return to. Earlier work, such as Hempel and Oppenheim (1948) or Braithwaite (1953), focused on simple, rather artificial, illustrations concerning the nature of explanation. These empiricist accounts were criticized via a series of by-now-famous hypothetical counterexamples: thought experiments concerning a flagpole's shadow or the mayor's Paresis (Salmon 1989, §2.3, provides a summary). Today's discussions of explanation have a different character. For instance, the large volume of 21st century work about mechanisms focuses on real-world cases from the life sciences. Likewise, recent interest in noncausal explanation (Reutlinger and Saatsi, 2018) is fueled by actual examples, drawn from scientific practice, and considering instances of mathematical explanation across different sciences in detail (Lange, 2017 illustrates this well).[3]

Our discussion concerns *evidential* uses of case studies and thought experiments: how these tools might justify (or undermine) philosophical theses. Both case studies and thought experiments play other roles, as we explain below. But insofar as evidence is concerned, we suggest that the near-elimination of thought experiments has been a wrong turn.

---

[2] Mizrahi (2020) provides quantitative evidence for the claim that case studies are an increasingly widespread form of evidence in the philosophy of science.

[3] Lange's 2013 paper on noncausal explanation uses a toy example, that of Mother and the strawberries. It is notable that this example is drawn from a 1972 paper (by David Braine.)

In part, our argument is grounded in an analogy between philosophy and science. We take our cue from the distinction between two central forms of empirical evidence: observation and experiment. Our argument doesn't require committing to either practice having better confirmatory prowess, nor do we propose that there is a perfect analogy between scientific and philosophical evidence. Rather, some of the limitations of observation, as well as some of the reasons why scientific experiments are so valuable, have important analogues in philosophical contexts. So, as we might think that experimentation and natural observation partially mitigate each other's evidential shortfalls, so we might think the same of thought experiments and case studies.

Case studies are not the only way to conduct philosophy of science in practice. Tools from social science such as ethnographies, structured interviews and surveys may be applied. So can formal methods, analytic models and simulations, and more data-driven approaches. These are all available, and increasingly used, by philosophers of science. We do not take a stand on these methods and as far as we can see our argument has no specific implications regarding them.

There seems to be a spectrum between handling instances of complex real-world science—case studies—and presenting a clean and simple hypothetical, as in a thought experiment. Some hypothetical cases might get complex or (more often) real cases might be treated in highly simplified ways: just as in the analogy with experiments and observations we lean on, there are borderline cases that do not fall clearly as one or the other. Indeed, some philosophical texts include variants on the same case, such that one variant seems more akin to a case study while the other more akin to a thought experiment (we discuss a relevant instance, drawn from work by Dellsén, below – section 6) But vagueness is not a conceptual concern here. There are many clear-cut instances at each end of the spectrum, and case studies are substantially more prominent relative to thought experiments, in many areas of current philosophy of science. Moreover, there is work in current philosophy of science in which case studies and thought experiments *are* combined, much in the ways we suggest. We can also be read as analyzing what such work gets right.

The paper is structured as follows. In Section 2 we consider what evidence in philosophical contexts is evidence for, suggesting norms for scientific rationality: what counts as a proper instance of confirmation, say, or as a successful explanation. In light of this, in section 3 we highlight some of the pitfalls of case studies. Section 4 sketches the analogy between thought experiments and case studies on the one hand, and experiments and observations on the other. In

section 5 we make good on the analogy, arguing that thought experiments can remedy many of the problems attending case studies. In the penultimate section, we consider potential criticisms. The most significant of these stems from recent work in experimental philosophy: while these represent serious challenges to the use of thought experiments, we suggest that the situation is not as bad in philosophy of science as it may be elsewhere.

## *2. What is (the Relevant Part of) Philosophy of Science About?*

Case studies and thought experiments are put to a plurality of uses in the philosophy of science. We focus on *evidential uses*. But evidence for *what*? We think most philosophers of science would accept the following broad answer: a central aim of the field is the attempt to characterize scientific rationality (or rationalities); by which we mean - attempts to say, in a detailed and precise way, what constitutes proper reasoning in scientific matters.[4] This includes questions about the nature of explanation, confirmation and classification, questions concerning the structure of theories and models, the role of probability and so on. It also includes attempts to characterize and individuate scientific kinds and categories, such as philosophical accounts of mechanisms or laws. We take for granted that scientific rationality is normative: it concerns *proper* reasoning.

As always, the normative is distinct from the descriptive. But perhaps unlike other normative endeavors, the philosophy of science typically operates under a presumption that science is successful. Indeed, this is a major motivation for philosophizing about science: it is seen as a paradigm of epistemic success, an important way of attaining knowledge and understanding. To be clear, this is a *presumption*, not a settled conclusion. Indeed, the numerous philosophers who doubt science's success—rightly taking a critical stance—usually think scientific work is epistemologically *assessable* in the sense that we can say whether instances of scientific reasoning are good or bad, successful or unsuccessful, and that the norms appealed to in making these judgments are not mere social conventions. And, of course, even if science as a whole is seen as a

[4] We take this to be a claim that unites monists, pluralists and particularists: they are all interested in proper scientific reasoning, they just disagree about the relevant scope and unity of that reasoning. Some parts of philosophy of science are fairly far removed from this project, and accordingly will not be covered. We have in mind primarily work that deals more directly with the questions that occupy science itself. Sometimes this is described as "philosophy of nature" (Godfrey-Smith, 2009), aiming to extract a big-picture, or to synthesize diverse scientific findings (Sterelny, 2003).

successful enterprise, many specific scientific areas and instances of research may not be so successful. Still, we think it is widely accepted among philosophers of science, even if implicitly, that science affords us a model of good reasoning and knowledge-attainment.

These remarks can make sense of the turn to practice: if philosophy of science is concerned with scientific rationality, in the normative sense, and if science is – at least often enough – epistemically successful, then philosophers of science are well-advised to look at how science works in order to discern the appropriate norms of scientific rationality. We look to practice because we presume that the ways science *actually* works are a guide to how it *should* work. Much of the work on mechanisms mentioned above has taken this path. Philosophers of science noticed the prevalence of mechanisms, especially in biology, and used this to argue for the (normative) significance of mechanistic explanation. Further, they appealed to how mechanistic explanations work in order to develop criteria for what should count as a good mechanistic explanation. (Craver and Tabery, 2019).

Philosophical claims made on the basis of case studies are typically intended to reach beyond the specific context of the case itself. Thus—simplifying somewhat—mechanists claim that in many parts of science a good explanation is one that uncovers underlying parts and their organization, i.e. the mechanism for the phenomenon. Authors then bring case studies, such as DNA replication (Darden, 1995) or the production of ATP (Bechtel and Abrahmsen, 2005), to buttress and refine this claim: the cases are supposed to show that, in (the relevant parts of) real-life science, explanations take a mechanistic form, and to further spell out how such explanations work. One way to view this is as an instance of inductive confirmation, much like in other empirical inquiries. One could also think of the role of cases in a less direct fashion (Schindler 2013; Schindler & Scholl 2020), or as more closely aligned with historical case studies (Kinzel, 2015). Either way, scientific success and case studies are connected via an evidential relation. Although we've focused on confirmatory cases here, case studies are also often used to refute philosophical claims. In the mechanism literature, for instance, Currie (2014) has argued that mechanistic models cannot accommodate some forms of narrative explanation on the basis of detailed accounts of explanations of Sauropod gigantism.

## *3. Problems with Case Studies*

So far, we've identified case studies as detailed examinations of particular episodes of scientific practice, intended to evidence claims about scientific rationality. Like a thought experiment, then, we're interpreting case studies as providing some kind of evidence for or against some philosophical claim. In this section, we survey some pitfalls and dangers associated with case studies, thus construed. We will be relatively brief, since our core claims are not new[5] But let us be very clear about our goals: we wish to prepare the ground for arguing that thought experiments can and should *complement* case studies. We are not arguing against the use of case studies; certainly not for their elimination.

***3.1 Cherry-picking***. Perhaps the most obvious and widely discussed problem with case studies is the potential for cherry-picking, i.e. the selective use of cases that suit a philosopher's thesis (Kinzel 2015, Schikore 2011, Currie 2015). Picking one's case to suit the argument is okay in some contexts – e.g. when one seeks to illustrate a claim rather than support it. But it clearly has a deleterious, biasing, effect on the evidential force of a case.[6] For if a case study is intended serve as inductive grounds for a more general claim about scientific rationality, then a good case study should be *representative* of relevant scientific work, that is, it should work similarly to a good inductive specimen.

The downsides of cherry-picking in this context aren't confined to direct worries about evidence. Case studies also shape debate around them: just as an influential thought experiment can center philosophical discussion so also can an influential case study become a kind of 'go-to' for thinking about particular philosophical issues. Some examples are the Michelson-Morley experiment for thinking about crucial experimental power, the Broad-Street Pump for thinking about medical testing, and the extinction of dinosaurs for thinking about historical inference. Insofar as these cases are not representative, cherry-picking can bias not only particular arguments, it has the potential to lead whole discussions astray.

---

[5] Apart from those listed below, precedents include Pitt (2001) and Lennox (2001).

[6] In principle, this problem might be mitigated by means of a more systematic meta-analysis, and appropriate sampling (Fuast & Mheel 1992, 2002). But such an approach faces obvious practical difficulties. Moreover, we suspect that it would still be subject to a version of the interpretive worry we raise below.

***3.2 Implicit philosophy.***[7] The majority of scientists are neither trained in, nor particularly interested in, philosophical analysis. However, scientific work is often influenced and shaped by scientists' own philosophical conceptions of what they are doing. This implies that in relying on case studies, philosophers might indirectly rely on accompanying philosophical judgements made by scientists – i.e. judgements concerning a philosophical question over and above (or generalizing from) the first-order scientific issue at hand. Thus, a scientist may judge a given model as better explaining some set of phenomena, but such judgements are (tacitly or otherwise) accompanied by more general claims about explanation. These latter judgments may at times be interesting, but often do not, nor do scientists have the training and motivation to perform them. Thus, scientists' presentations of their work sometimes embody implicit philosophical content. Moreover, such implicit philosophical all too often reflects a deficit of serious thought and consideration. Because it is not their main interest, nor what thy are trained for, scientists often invoke agenda-driven, potentially careless, philosophical judgements – made as a "side-kick" to the first-order, scientific task. In using case studies, then, we are liable to take on these unreflective implicit philosophical ideas. The problem here isn't with the judgements coming from non-philosophers. The problem is it's the potential for careless or agenda-driven judgments, and ones that are often left implicit.

***3.3 Interpretation.*** A further source of bias in the study of cases concerns interpretation. It is widely accepted in the philosophy of history, and in methodological reflections amongst historians, anthropologists, and others focused on interpretive pursuits, that particular episodes are amenable to multiple interpretations. This holds equally for case studies and poses a distinct, and potentially deeper problem than cherry-picking. Any given case is amenable to multiple, evidentially divergent interpretations. One's chosen interpretation must therefore be defended as appropriate, and indeed as better than interpretations which act against their normative thesis.

For instance, given an instance of explanation, the philosopher must, *inter alia*, answer: What kind of explanation is present? Is it a good explanation? What is the explanation's structure? Such questions are crucial to the case's evidential value, but they are not self-evident, nor are they uniquely determined by the case *per se*. To be clear, the problem isn't that such interpretations are

[7] This point should, ideally, be evidenced (presumably by one or more fairly detailed case studies). We do not have space for that. So we posit it as a possibility which, if realized, poses a concern about the use of case studies.

needed, but that, on the one hand, they can be strongly affected by the interpreter's preexisting philosophical views. And on the other hand, the case study's audience doesn't, typically, know the case firsthand and is thus hard put to voice a critical viewpoint. Indeed, oftentimes the audience doesn't have necessary background knowledge or, for another reason, cannot access the case independently, and is thereby entirely dependent on the case study's originator to gain understanding of the case.

***3.4 Manipulation and isolation.*** Our last concern about case studies is simple but crucial: they cannot be manipulated. Often, the debate in philosophy of science (as in science itself) concerns the significance of one or a small number of conditions. One side argues, say, that a good explanation must describe the explanandum's mechanistic underpinnings, the other side says it needn't. A good test of such a claim would consider cases that, as far as possible, differ only in the presence of mechanistic detail (where the explananda are as similar as possible, both explanations are couched in similar formal terms etc.) But by their very nature, case studies do not allow for this. They are too complex, too detailed and too messy: the very idiosyncrasy of scientific practice that philosophers of science aim to take seriously in their work undermines clean comparisons. Relatedly, it is often hard, at times impossible, to abstract from cases as we might wish. More precisely: it is difficult to ignore details of the case that are less closely related to the philosophical point, while still discerning how the case behaves.

For this and related reasons, when considering the status of a case in actual science – the judgements of scientists about it, or its standing relative to other comparable results – one must settle for judgements about the case as it actually is. It is difficult, in the course of a "pure" case study, to cast judgement about some of a case's features and not others, or about how its status would change under counterfactual circumstances.

## *4. Experiments versus Observations*

Having outlined some concerns about case studies, we now present our positive proposal, arguing for the use of thought experiments. To do so, we first draw an analogy with evidential methods within science.

Both observations and experiments can provide support for, or weaken, a given hypothesis, and for basically the same reason: both are ways in which we, as investigators, make causal contact

with a phenomenon, ascertaining whether and how well it aligns with our hypothesis. Following Currie & Levy (2019)'s way of putting things, both, when successful, are ways for an investigator to study a *specimen* – an ideally typical instance of the phenomenon under investigation. It is specimenhood, when and to the extent that it obtains, which allows results concerning a particular object to generalize to other members of the relevant kind. Here we will not delve more deeply into specimens and their epistemic properties.[8] Instead, we're interested in the relative strengths and weaknesses of observations and experiments, with an eye to parallels with case studies and thought experiments. In a nutshell, the difference concerns the immediacy of observation versus the manipulability of experiment. Let us explain.

In an *experiment*, we bring a specimen into the lab, i.e., into a set of contrived conditions, and conduct a controlled manipulation of it. In contrast, an *observation* involves going out to the field (or sending probes to the ocean floor, looking through a telescope, etc.) and collecting data about a relevant portion of the world. If said portion is indeed relevant, and in particular if it can be taken to be a specimen, then the observations may well be suitably generalized. However, unlike in an experiment, in observation researchers are relatively passive – the specimen is not intervened upon to the extent that we would in the lab. Naturally, there are many gray areas and intermediate cases. For our purposes, what matters is the difference between what we take to be experiments and natural observations understood as aligning insofar as both involve specimenhood, but diverging insofar as forms of manipulation are involved. Observation is more passive but less mediate, and vice versa for experiments.

A key benefit of experiments lies in the capacity for manipulation. Relying on work by Okasha (2011), this can be fleshed out as follows. Suppose, for simplicity, that we are testing a hypothesis of the form "all Fs are G".[9] Here, the most relevant test would take the conditional form $Fa \rightarrow Ga$, i.e., cases where we bring about Fness and try to detect Gness. In particular, these observations would be preferable to evidence from conjunctions, of the form $Fa \,\&\, Ga$. The former is a *bona fide* instance of the hypothesis, and hence of direct relevance to its truth. In contrast, the

[8] That said, see Guala (2002) and Morgan (2005) for emphasis on the materiality of experiments, and Parke (2014) for some pertinent concerns about this.

[9] We understand such conditionals as having being stronger than material implication. They have modal, typically explanatory import i.e. they state that G in some robust sense depends on F, or that F explains G. We suspect, but aren't sure, that Okasha views the matter similarly.

latter, conjunctive claim, needn't be an instance of the generalization in question (it may, e.g., be an instance of 'all Gs are F'). Hence it has less confirmatory relevance. Put differently, when exerting control in the context of an experiment, one's object of study is isolated, the condition of interest is bought about, and we see whether some consequence (co-)occurs. In contrast, in performing a non-experimental (i.e. non-controlled) observation, one merely observes the two conditions jointly occurring, which is of lesser evidential significance with regards to the generalization at issue.

Moreover, in an experiment one can manipulate the specimen repeatedly. This allows the generation of larger samples, thus providing firmer data. It also allows one to vary the precise manipulation undertaken, often in a fine-grained manner enabling systematic understanding of how the specimen behaves under subtly varying conditions. Both features, and especially their co-occurrence, aren't typically present in observation.[10]

It is crucial to note that the distinction doesn't only involve continua, but also will likely play out in very different ways across different contexts in which scientists work. In studies of animal cognition and behavior, comparative psychologists test animal capacities experimentally in labs whilst ethologists attempt to observe them in the wild. In medicine, epidemiological studies are observational, relying as they do on the collection of data of rates of infection in human populations, while clinical trials are experimental insofar as they involve interventions on controlled groups. However, even across these examples there are differences in what counts as a typical specimen, our capacities to make interventions, and so on. Thus, the distinctione applies differently across the differing situations scientists work within. The model is general in a sense: provide an example of data-gathering from a set of specimens and we can understand it is more experimental or more observational (or more experimental or more observational given some contrast) on the basis of the extent to which it involves isolation and manipulation, understood as relevant to that context. As such, we don't see Currie & Levy's framework as obviously in tension with more historicized accounts, which are often (perhaps rightly) skeptical of distinctions between 'experiment' and 'observation' as understood in less context-sensitive ways (see for instance Dustan & Lunbeck 2010, Malik 2017).

[10] We set aside so-called "natural experiments", which are a form of observation sometimes regarded as evidentially on a par with experiments (e.g. Woodward 2003, 94). We do not regard this way of thinking of "natural experiments" as fully satisfactory, but the issue is of minor relevance here.

Crucial to our analogy, Currie &Levy think there is a tradeoff of sorts between the evidential value of more experimental and more observational features, within contexts. While an experiment provides fine-grained information, through repeatable control, it is also harder to generalize from it, since the conditions under which data were obtained are typically different, to some extent, from natural conditions under which the phenomenon typically occurs. In contrast, an observation involves little or no alteration of conditions, and so gives one a more direct "look" at the phenomenon, and more readily generalizable outputs. There is no doubt something to the sense that observations are less intrusive, relative to experiments, and therefore epistemically "purer", but we musn't exaggerate the point. An observation too is an instance of scientific agency, involving a choice of what to observe and by which means. In an observation, as in an experiment, one must make decisions about which data to include in one's analysis, how to classify data etc. And while it matters that observations do not involve intervention, this does not make them into a form of unmediated contact with nature.

So, an argument could be made that these features give experimentation a confirmatory edge: an experiment provides stronger evidence than a corresponding observation. In response, one might emphasize the artificiality of experiments: that specimens are manipulated *by us* in a non-natural setting introduces confounders and limits their representativeness. Our claim doesn't turn on one approach or the other winning an epistemic boxing-match, however: as we'll claim for thought experiments and case studies, paradigm natural observations and experiments are complementary. Each method has distinctive features; they can and should be put to use in tandem. This, of course, is true for scientific cases as well: in principle, at least, evidence from epidemiological studies and from clinical trials could be brought to bear on questions of disease, as results from comparative psychology and ethology could inform us about cognition in non-human animals.

### *5. A Plea for Thought Experiments*

The shape of our position should by now be clear: thought experiments are to the philosophy of science what experiments are to science itself. Before stating the argument in detail, we make a few remarks about thought experiments.

***5.1 Thought experiments in the philosophy of science***. There hasn't been much (if any) discussion of thought experiments in the philosophy of science *per se*. There has been a fair amount of discussion of the nature of thought experiments in philosophy generally (e.g., Sorensen, 1992b; Häggqvist 1998; Williamson, 2004, 2008; Machery, 2017 Ch. 1) and in science (Norton, 2004; Gendler, 2000, 2004; Stuart 2016; Brown & Fehige, 2019) and some comparative discussions, too (Horowitz and Massey, 1991; Stuart et. al. 2018 Murphy 2021, Cooper 2005). Because we argue for a mixed-methods approach to the philosophy of science—that is, we're interested in the combination, and in this particular context—these discussions are of limited help here. We will note, however that the main distinguishing mark of our discussion is not a claim about what thought experiments are or how they work. We accept many of the claims made by the authors just noted e.g., the idea that a thought experiment involves the imagination and that they resemble, and to some extent overlap with, counterfactual reasoning. Our claim, however, is about the evidential role of thought experiment, as this pertains to normative issues in philosophy of science. We do not think this implies, and certainly aren't committed to, the claim that thought experiment have evidential power in other domains or uses.

We take thought experiments to be depictions of hypothetical scenarios, or contrived examples, typically abstract and simple in character. Famous counterexamples to the DN model are a good case in point.Here is one, taken from the work of Sylvan Bromberger[11]:

"There is a point on Fifth Avenue, M feet away from the base of the Empire State Building, at which a ray of light coming from the tip of the building makes an angle of Θ degrees with a line to the base of the building. From the laws of geometric optics, together with the "antecedent" condition that the distance is M feet, the angle Θ degrees, it is possible to deduce that the Empire State Building has a height of H feet. Any high-school student could set up the deduction given actual numerical values. By doing so, he would not, however, have explained why the Empire State Building has a height of H feet, nor would he have answered the question "Why does the Empire State Building have a height of H feet?" nor would an exposition of the deduction be the

[11] As can be seen, this thought experiment is essentially identical to the better-known flagpole case. Interestingly, the flagpole version was never presented by Bromberger in print. It was made famous by Hempel, to whom Bromberger presented the flagpole case in person. (See Bromberger, 1992, 8).

explanation of or answer to (either implicitly or explicitly) why the Empire State Building has a height of H feet" (Bromberger 1966, 92)

No one (no high schooler, in particular) has, so far as we know, ever actually "explained" the height of the Empire State Building in this way. But we can readily understand such cases and make judgements about them (in this case, judgments about their explanatory power). What matters here is not the case's hypotheticality *per se* but rather its *constructed* nature. The philosopher can set up scenarios as she pleases, depending on what she finds interesting, or what she thinks will have significant probative force. In particular, one can perform something analogous to a controlled experiment – where all relevant variables are fixed except for a factor of interest. Indeed, it is possible, in nearly every case, to do this for many, if not all, factors of interest. In cases like the building and the shadow, one can vary the explanans, the explanandum, the context and so on (Indeed van Fraassen [1980, Ch. 5] does so.)

Let's see how this plays out in the Empire State Building case. The logical empiricist account identified explanation with derivable prediction: a genuine *explanans* must include a natural law, and their conjunction must logically imply the *explanandum*. The Empire State Building thought experiment leads us to focus on the *symmetry* of these accounts. As the length of the shadow may be derived from the length of the building, and pertinent assumptions about the behavior of light, the DN account implies that the building's height explains the shadow. However, as the length of the building may be derived from the length of the shadow (plus the behavior of light), the DN account implies that, symmetrically, the shadow explains the building. This stands in tension with a normative judgment elicited by the case: the building's height is rightly seen as explaining the shadow, but not vice-versa. The immediate lesson is that prediction via derivation might be symmetrical, but explanation is not; thus, the two shouldn't be equated. Note that, in this case—and in some other, though not necessarily all, cases—there was also a further lesson: that future views of explanation must account for explanation's asymmetry. This proved to be an important motivation for causal views of explanation, which took center stage in the wake of the DN account (Levy forthcoming, Chps 2-3.) Thus, the Empire State Building case doesn't only play a negative role, but a positive one as well: supporting asymmetric accounts of explanation such as causal ones.

We can interpret this episode – which we view as a near best-case scenario for the use of thought experiments in the philosophy of science – as follows. The thought experiment provides us

with an artificial specimen of an explanation. This simple specimen isolates the phenomenon of interest, namely, the putative asymmetry of explanatory relationships. Note further that this specimen is easily manipulated: we can imagine variations on the thought experiment, thus exploring how our judgements about explanatoriness shift with differences in the case. The process of isolation and manipulation teaches us about the sensitivity of our judgements of explanatoriness, and, on that basis, might be taken as reason against the DN view, and to for asymmetric models such as causal accounts. In this sense, at least, the *experimental* aspect of thought experiments should be taken very seriously.

We should note that our emphasis on the experimental side of thought experiments is compatible with different views on their epistemic function. For example, discussing thought experiments in science, Norton (1996, 2004) argues that they are best understood as arguments—as opposed to experiments—and both Buzzoni (2008) and Stuart (2016) respond by emphasizing their experimental aspects. Epistemically deflationary accounts like Norton's (and more ambitious ones like Sorensen, 1992a) need not deny that thought experiments, like 'true' experiments, involve systematic manipulation and exploration. They rather say that manipulation and exploration are ways of constructing argumentative variants.[12]

There is another way in which thought experiments differ from "real" experiments: simplicity. In a real-life experiment our specimen may be simpler than its "wild" counterpart, but is nonetheless complex insofar as it is a concrete system, with a complex internal structure and embedded in a multitude of interactions with its environment. In contrast, a thought experiment can often be devised so as to be truly simple – as a creature of our imaginations, it can have all and only the features of interest to us. In this sense, as we argue below, thought experiments enjoy an advantage over case studies that has no clear parallel in the within-science distinction between

[12] Norton's position might be thought to imply a different concern for us, namely the thought experiments cannot provide *evidence*, inasmuch as they "are merely picturesque argumentation" (Norton, 2004.) Our underlying conception of thought experimentation appears—we are not entirely sure—to differ from Norton's, inasmuch as he takes them not to involve judgements. We cannot here enter into an extended discussion of these points, so we will merely comment that we view Norton's account as less plausible in the context pf philosophical thought experiments (relative to scientific ones). Conversely, we think that while thought experiments in philosophy are better seen as probes of our judgement tendencies, and reconstructing arguments on their basis is important, this does not quite exhaust their content, as a Norton-style view would hold.

experiments and observations (although is likely available via other methods: modeling most obviously).

***5.2 How thought experiments mitigate case studies***. Let's now look at how the pitfalls of case studies considered above can be mitigated by thought experiments.

First, we worried that scientists' unreflective second-order ideas about science might be smuggled into an analysis via a case study. Thought experiments are specifically devised by philosophers to be put towards particular tasks. Their construction serves to lay bare and make explicit and transparent the second-order claims that might otherwise have lain hidden: we see exactly this in how the Empire State Building case highlights the symmetry (or otherwise!) of explanatory relations. Here both the simplicity and the manipulability of thought experiments is relevant: given simplicity, we can be more confident that we have identified the relevant features, and that our judgements are based on features that we can identify. Given manipulability, we can be more confident that the features that are relevant have been controlled for and varied in ways that would make them distinguishable.

Next, cherry-picking. Here the problem is along the lines of: you claim that this is a representative specimen of successful science, but you've grabbed a case that too-suspiciously fits your account, without giving me any independent reason to view it as representative. Thought experiments cannot, of course, transform a cherry-picked case into a representative one. But they can help us see whether the case we have picked has features that we'd expect. In particular, it can provide a normative explanation as to why we should expect the case study to be so. Here, we see how a case study and a thought experiment might work together to evidence a claim about scientific rationality. As in science, our credence in philosophical positions is set, in part, by background theories concerning what we expect of patterns of reasoning. Thought experiments are one way of testing and evidencing those background theories, and thus together with a case study can evidence a particular normative claim.

A further, related advantage of thought experiment in this context has to do with their 'cost'. If it appears that a philosopher brings up a hypothetical case in a manner that suits their particular theoretical aims, one can always present a different hypothetical. Unlike case studies, thought experiments are "cheap" – they require less time and expertise to put together (here indeed is another disanalogy with concrete experiments.) Indeed, debates over the DN model involved a

multiplicity of cases, some (primarily examples from Hempel himself) telling for and most (including examples reviewed in Salmon, 1989 2.3, as well as cases of narrative causal explanation by Scriven, 1962) telling against it. It was only given this multiplicity of cases (as well as other types of arguments, of course) that the DN model fell from grace.

Third, interpretive difficulties do not vanish, but diminish. As we saw, particular cases are amenable to multiple interpretations, and as differing interpretations might support different philosophical claims, we require independent justification of those interpretations. As in cherry-picking, we can see that thought experiments both can be used to provide that independent justification and are themselves less amenable to interpretive worries. Philosophical interpretations are normative, and thought experiments motivate and adjudicate normative judgments. As such, thought experiments like the flagpole case can be used to defend interpretations of case studies that appeal to asymmetric explanatory relations against those positing symmetrical ones. Just as with experiments and observations, one does not *a priori* trump the other. Rather, they play different roles: the case study provides a richly textured example of scientific success, the thought experiment isolates and evidences the normative principles (potentially, those suggested by the case.)

Thought experiments themselves are, of course, open to worries of interpretation, but we suggest that they're comparatively transparent. In particular, thought experiments typically do not require training or special familiarity with some domain of research. In this sense, they are more accessible and what makes them tick is more apparent. Further, the simplification and manipulation afforded by thought experiments enable differing interpretations to be systematically examined. It would be an exaggeration to say that thought experiments wear their assumptions on their sleeves – but they do, we think, wear them closer to the sleeve relative to case studies.

Finally, in constructing a thought experiment, we are free to abstract as we please. Moreover, we can manipulate a (pre-existing) thought experiment when the need arises. Thus, the concern that case studies do not allow us to differentiate between relevant factors and test their significance is mitigated. We can move beyond mere conjunctions and test the conditional we are after; and we can do this repeatedly, looking at variants of the example, as needed. As we have suggested, the proliferation of cases that brought down the DN model – and no less importantly, motivated causal

approaches – can be read in this way, i.e. as experimental variations. They serve as conceptual replications, as it were.

All in all, it seems that thought experiments can counterbalance a case-based methodology. We've suggested they do this in two ways. First, thought experiments are not subject to the same worries that case studies are. Their manipulability, simplicity and transparency allow them to identify, isolate and critique otherwise unreflective second-order claims, cherry-picked examples and concerns about interpretation. Second, thought experiments and case studies complement one another. Where case studies ensure that the rubber hits the road in our thinking about science—that we do not abstract so far that we lose touch with the features that in fact lead to scientific success—thought experiments allow us to isolate and manipulate the normative principles that play out in those case studies. They are complementary methodologies and philosophers are best advised to use both.

## *6. Problems with Thought Experiments?*

We've argued that thought experiments can and should play an evidentiary role in philosophy of science. Hopefully it is clear that we think a philosophy of science relying on thought experiments alone would be impoverished: they precisely lack the rich detail of practice which case studies afford, and which is often necessary to do science philosophical justice. But there is still space to object to our claim that thought experiments deserve *any kind of place* in the philosophy of science, even merely the complementary one we've suggested.

One kind of concern might turn on a general resistance to a priori means, according to which they lack precisely those features that make "real" experiments valuable, thus vitiating the analogy we drew between evidence in science and in its philosophy. Another problem stems from empirical results suggesting that thought experiments are prone to artifacts and biases. In this penultimate section, we attempt to diffuse these concerns, while also expanding on how we view the interaction between thought experiments and case studies.

***6.1 What are we experimenting on?*** A critic of our argument may appeal to two potential disanalogies between thought experiments and scientific experiments. First, it is unclear whether, in thought experiments, the phenomena probed are sufficiently stable and robust. Perhaps they can give us a glimpse of our concepts and conceptions, but what ensures that thought experiments

supply insight into genuine principles of scientific rationality? In other words: what ensures that they allow us to make contact with the relevant phenomena – in this case, cannons of prediction, explanation and other facets of scientific practice? Second, results of "ordinary" experiments can be re-confirmed and experimental instruments re-calibrated. This way their reliability can be assessed and improved over time. What ensures reliability in gedanken experimentation?

To some extent, our suggestion that thought experiments complement case studies, rather than replace them, should speak to these worries. If indeed a philosopher was happy to rest on their gedanken laurels, and not work to show that their thought-experiment-derived principles actually played out in successful science, that would be cause for concern. However, unlike contexts in which armchair means are the sole sources of evidence, case studies can be seen to supply a counter-balancing perspective, through which one can check and in effect (partially) calibrate the results of thought experiments. The lessons one draws from a "clean" thought experiment can thereby be calibrated, as it were, against the "messier" realities reflected in case studies. While, admittedly, this is not a direct and powerful way of establishing external validity (Wilson, 2014), it is a significant step in that direction. This also suggests one form an investigation combining cases and thought experiments might take: one isolates a candidate principle by means of thought experiment, and then looks to case studies as a means to corroborate and flesh out the principle. In some ways, this was the trajectory followed by work on explanation we appealed to earlier: thought experiments to the DN model, including (but not restricted to) Empire State type cases, served as more than counterexamples. They also pointed to the fundamental asymmetry of explanation, and hinted at that causation was key to explanation. Further work proceeded in part by appeal to real-life cases, in light of which the causal approach was both vindicated and fleshed out.[13]

Other approaches might turn this on its head: scientific practice could suggest principles which are explored and tested via thought experimentation. As a relatively minimal, yet still instructive case in point, consider an example from a recent paper by Dellsén (2016). Dellsén discusses the question of whether scientific progress is aimed at knowledge or at understanding. To support the latter option, he presents a (brief) case study: Einstein's explanation of Brownian

[13] Note that we do not aim to suggest that this the only way of combining thought experimentation and case studies. Indeed, below we comment on the mechanistic approach, the development of which did not quite follow this route.

motion in one of his famous annus mirabilis papers (Einstein, 1905/1956). Dellsén presumes that this is an undisputed case of scientific progress. But it is a case in which, he argues, Einstein had very partial information about the phenomenon he was accounting for, and hence could not even be sure it was an explanation of Brownian motion. So he could not have had knowledge of the explanation – obviating the idea that progress is a matter of attaining knowledge. Dellsén then considers whether someone might dispute the idea that the kinetic theory of heat, and the phenomenon of Brownian motion, were in fact unknown at the time of Einstein's work. To this end he constract a thought experiment in which "Einstein's explanation was put forward before the kinetic theory of heat became sufficiently justified to be known…" But he argues that "[n]one of this would take away from Einstein's achievement, which was to show how Brownian motion is explained by the kinetic theory of heat ." (ibid., 76). Thus Dellsén manipulates, in the imagination, a case that originates in actual science.

Further, we doubt the sheer a priori-ness of thought experiments. As noted earlier, we take it that philosophy of science is concerned, at least in part, with articulating norms of scientific rationality in a manner that both explains, but also forms the basis for a potential critique of, scientific practice. As such, we take the concepts and the principles they are involved in – the primary "matter" upon which thought experiments operate – to be formed in a manner that is sensitive to the practice of science (including its successes and failures.) This is not the context to elaborate on a theory of concepts and concept formation. But let us note that broadly empiricist theories of concept formation do allow, if developed in a suitable way, for a view of thought experiments that regards them as indirectly probing what lays outside of the mind as such. A recent example, applied directly to problems in philosophical methodology, is Strevens (2019). In this way one can view thought experimentation as both a priori, in one sense of the term, but also as informed by practice and sensitive to it.

So, our discussion of thought experiments does not take them to be *a priori* in a pernicious way because they do not work alone, but in tandem with case studies. Recall that case-study oriented work sometimes relies on unexamined normative principles such as scientists' unreflective second-order judgements, which could underwrite both which cases we pick and how we interpret them. Thought experiments—at the very least—serve to make those underlying principles more explicit and open to critique. We can imagine a continuum from, on one end, someone who thinks thought experiments have no evidential value and cannons of rationality can only be gleaned

directly from scientific practice. On the other end, someone might think case studies are purely descriptive (whatever that means) taking their normative force to come from judgements elicited by thought experiments. Regarding the former end of the continuum, we've argued that thought experiments can play crucial supporting roles by mitigating the potential pitfalls of case studies, and this holds even if ultimate justification lies in practice. For the latter extreme, case studies are still important inasmuch as they allow us to flesh out principles and show how they work. So, even on the two extreme ends of the continuum we have good grounds for caring about both case studies and thought experiments. We presume that most philosophers would locate themselves somewhere in between these two extremes. There is plenty of space for productive and interesting methodological disagreement within that space.

***6.2 Worries from experimental philosophy?*** The so-called 'negative program' in experimental philosophy has revealed potential biases and unreliability in judgements stemming from thought experiments. (Knobe and Nichols, 2017). The focus has been on ethics, philosophy of language and epistemology—not thought experiments in the philosophy of science—and so we'll adapt that criticism to the present context. Two main problems seem relevant here.

Problem One: there is demographic and cultural diversity in responses to thought experiments. Some thought experiments elicit different responses in different people, for instance between Eastern and Western subjects (Weinberg et al., 2001, Machery et al., 2004) or between genders (Buckwalter & Stich, 2013). This has led some to object that either philosophical analysis fails to capture the breadth of conceptual diversity across cultures, or that the use of thought experiments must privilege some demographic(s), perhaps especially Western philosophers. This assumption lacks justification, the argument goes, reflecting no more than a parochialist attitude or a kind of Dogmatism (Machery 2017, Ch. 4).

Our response is twofold. For one thing, we reiterate that our recommendation is to combine thought experiments with case studies. This should help keep biases—in either thought experiments or, of course, case studies— at least somewhat in check.[14] Moreover, the response to problems of diversity in thought experimentation should partly take the form of exploring a wide range of cases,

[14] We should note that insofar as science itself reflects gendered, western and other biases, there are real limitations to this mitigation strategy

from across different sciences, and perhaps different cultural contexts (and/or different scientific cultures.) And the construction and use of thought experimentation should also be sensitive to this issue, in at least two ways. First, one should attempt to elicit judgements from a variety of audiences across different demographic contexts, including from those working in different scientific contexts. That is, thought experiments need not be fodder just for philosophical consumption but, in line with experimental philosophy's positive program, something we put to scientists—and a diversity of them. Second, one should not suppose that principles generated from, or tested by means of, thought experimentation are either applicable in a very general manner, and/or that they will have a uniform (or uncomplicated) applicability across scientific practices. The literature on mechanistic explanation, can, to an extent, be understood this way. It was given initial impetus by the move towards the causal approach that was buttressed (as we discussed above) by thought experimentation. But by carefully studying a diversity of cases from across different parts of biology and the social sciences) philosophers have come to appreciate the different sorts of mechanisms, and the different normative principles governing mechanistic explanation.

But we also think that, as regard diversity of responses to thought experiments, the situation in philosophy of science is somewhat different than in other areas. In particular, the idea of privileged judgements seems more plausible. What we have in mind is this: those well-acquainted with science are to be expected to have sounder and more precise judgements, inasmuch as these judgements reflect the accumulated wisdom, so to speak, acquired by a keen observer of the sciences (or of a particular science, as the case may be). Here, too, then, a case for the importance of familiarity with the practice emerges, even though it does not manifest itself in a direct discussion of episodes from real-world science. Differently put: a person who has taken a close look at science, or some part of it, is better placed to make judgments about its workings, including conceptual judgements pertaining to categories such as explanation, prediction and the like.

This latter defense is similar to, but not the same as, what is sometimes called 'the expertise defense' in discussions of thought experimentation and experimental philosophy (e.g. Williamson, 2008, Ch. 6). The idea is that even though thought experiments in ethics, say, exhibit differences across cultures or demographics, these should not worry us because it is only the judgements of experts, professional ethicists in this case, that we should rely upon. Typically, the expertise defense rests on the idea that philosophers have well-honed conceptual skills, analytic abilities and related competencies (and it is often further argued that the case is akin to science, where only a trained

expert's findings are taken seriously.) This is not our suggestion. Rather, when a person pays close attention to a certain practice, this gets reflected—positively so—in her response to thought experiments. The basis for sound judgements in thought experiments is a function, in part, of our acquaintance with how science works. It is just that fleshing out the acquaintance by means of thought experiments lends it evidential strengths that other forms of reliance on practice typically lack.

Next, let us tackle Problem Two: presentational effects. Work in experimental philosophy suggests that seemingly irrelevant aspects of the manner in which thought experiments are presented can affect the judgments elicited. For instance, several studies have shown that the order in which thought experiments are presented to subjects can make a marked difference to their response to scenarios (e.g. Wright 2010). Likewise, when subjects are induced to experience irrelevant emotions, this may affect those judgements (Greene, 2015).

We agree that, in general, presentational effects are worrisome. We should guard against them, just as we should if the results of concrete, scientific, experiments vary according to when they were obtained or the emotional state of the lab's personnel. But we think that such concerns should not lead us to abandon thought experimentation and forgo its fruits. First, let us remind the reader, yet again, that we recommend a *combination* of thought experimentation and case-based argumentation. As noted above, this can mitigate some of the concerns about thought experiments, and we believe it can help rectify errors of a presentational source, too. Similarly to how scientists take steps to control for artifacts from subjectivity (take steps, but likely never totally eradicate them!), so too should philosophers control for such artifacts. And, indeed, case studies are a brilliant tool for just this.

But the more important point in responding to presentational issues is that thought experiments, as emphasized in the discussion of "regular" experiments, can be repeated and varied. As we noted in discussing thought experiments concerning scientific explanation, part of the reason for viewing this as a success story for the use of thought experiments is the fact that the debate surrounding the DN model involved *multiple thought experiments*, probing the issues from various angles, as well as *variations on the same thought experiment.* Our account of the benefits of thought experimentation suggests that this is an important aspect of the proper application of this method. Our suggestion should be understood with this mind. In other words, we recommend that thought

experiments be used in a systematic and serious way: we should (as a community, at least, if not as individuals) aim to produce a variety of thought experiments, isolating factors in a systematic way. We should attempt to vary as many of the factors that are supposedly playing a role in shaping our judgements across a range of scenarios. We can aim to get feedback on thought experiments from people in multiple demographics And, in some cases at least, we should to test for presentational and other artefacts, if necessary in the lab.

More generally, we fully accept that philosophers should pay attention to results in psychology and experimental philosophy and learn from them where potential artifacts lie. We take seriously the *experimental* in thought experiments, and just as information about potential confounders is crucial for proper experimental design in science, so should lessons from experimental philosophy aid the design of thought experiments.

## *7. Conclusion*

The rise of case studies in the philosophy of science was closely linked with post-positivist, historically-minded thinkers and led to decidedly practice-based forms of work. These approaches bore many significant fruits, leading to a philosophy of science that is truer to real-life science and attentive to the diversity inherent in the practice of science. But the practice turn also had significant costs: it led to a fragmentation, both within philosophy of science and in terms of its connection with other areas of philosophy. It has also, we suggest, led to a philosophy of science that rests on a poorer evidential basis than it could.

We have argued that an appeal to thought experiments, especially if it augments, rather than replaces, case studies, can help remedy this situation. Case studies are prone to cherry-picking and biased interpretation; they are hard to treat abstractly and to repeat and manipulate. Thought experiments are, in some important respects, a mirror image: they are "cheap" and hence can be repeated and manipulated, partially mitigating these problems.

Questions naturally arise as to how to combine thought experimentation with the study of cases. We have offered some suggestions on this score, but we are hesitant to make definite recommendations. Surely, there is no single recipe or method to go by. One option, which we briefly described at in the previous section, is to start with a real-life case and modify it in thought.

Another is to apply case studies and thought experiments at different levels of abstraction. There are several other possibilities. We would consider it a step in the right direction if such approaches, or their relatives, were adopted more frequently in philosophical discussions of science.

***Bibliography***


Ankeny, R., Chang, H., Boumans, M., & Boon, M. (2011). Introduction: philosophy of science in practice. *European journal for philosophy of science*, 1(3), 303-307.

Bechtel, W. and Abrahmsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*. 36 (2):421-441

Braine, David. 1972. "Varieties of Necessity." *Supplementary Proceedings of the Aristotelian*

*Society* 46: 139– 170.

Braithwaite, R. (1953) *Scientific explanation. A Study of the Function of Theory, Probability and Law in Science*. Cambridge University Press.

Buckwalter, W. and Stich, S. (2013). Gender and Philosophical Intuition. In Knobe, J. & Nichols, S. (eds.), *Experimental Philosophy*, Vol.2. Oxford University Press.

Buzzoni M (2008) Thought experiment in the natural sciences. Königshausen & Neumann, Würzburg

Cooper, R. (2005). Thought experiments. *Metaphilosophy*, 36(3), 328-347.

Currie, A. (2015). Philosophy of Science and the Curse of the Case Study. In *The Palgrave handbook of philosophical methods.* Palgrave Macmillan.

Currie, A. M. (2014). Narratives, mechanisms and progress in historical science. Synthese, 191(6), 1163-1183.

Currie, A., & Levy, A. (2019). Why experiments matter. *Inquiry*, 62(9-10), 1066-1090.

Daston, L., & Lunbeck, E. (Eds.). (2010). Histories of scientific observation. University of Chicago Press.

Darden, L. (1995). Exemplars, Abstractions, and Anomalies: Representations and Theory Change in Mendelian and Molecular Genetics," in Lennox, J. and Wolters, G. (eds.), Concepts, *Theories, and Rationality in the Biological Sciences*. University of Pittsburgh Press

Godfrey-Smith, P. (2009). *Darwinian Populations and Natural Selection.* Oxford University Press.

Guala, F. (2002). Models, simulations, and experiments. In Model-based reasoning (pp. 59-74). Springer, Boston, MA.

Häggqvist, S. (1996). *Thought Experiments in Philosophy*, Almqvist & Wiksell International.

Hempel, C.G. and Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15(2): 135-175.

Horowitz, T., and Gerald M. (eds.), 1991, Thought Experiments in Science and Philosophy, Lanham: Rowman & Littlefield

Kinzel, K. (2015). Narrative and evidence. How can case studies from the history of science support claims in the philosophy of science?. *Studies in History and Philosophy of Science Part A*, 49, 48-57.

Knobe, J. and Nichols, S. (2017). Experimental Philosophy. *Experimental Philosophy", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/>*.

Lange, M. (2017). *Because Without Cause: Non-causal Explanations in Science and Mathematics*. Oxford University Press.

Lennox, J. G. (2001). History and philosophy of science: A phylogenetic approach. *História, Ciências, Saúde-Manguinhos*, 8(3), 655–669.

Machery, E. (2017). *Philosophy within Its Proper Bounds*. Oxford University Press.

Malik, S. (2017). Observation versus experiment: an adequate framework for analysing scientific experimentation? Journal for General Philosophy of Science, 48, 71-95.

Mizrahi, M. (2020). The Case Study Method in Philosophy of Science: An Empirical Study. Perspectives on Science 28 (1): 63–88.

Morgan, M. S. (2005). Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology*, 12(2), 317-329.

Murphy, A. (2020). The aesthetic and literary qualities of Scientific Thought experiments. *In The Aesthetics of Science* (pp. 146-166). Routledge.

Norton, J. (1996) Are thought experiments just what you thought? *Canadian Journal of Philosophy* 26:333–366.

Norton, J. (2004). On Thought Experiments: Is There More to the Argument? *Philosophy of Science*, 71: 1139–115

Okasha, S. (2011). Experiment, Observation and the Confirmation of Laws. *Analysis* 71 (2):222-232

Parke, E. C. (2014). Experiments, simulations, and epistemic privilege. *Philosophy of Science*, 81(4), 516-536.

Pitt, J. C. (2001). The dilemma of case studies: Toward a Heraclitian philosophy of science. *Perspectives on Science*, 9(4), 373–382.

Reutlinger, A. and Saatsi, J. (eds.), (2018). *Explanation Beyond Causation; Philosophical Perspectives on Non-Causal Explanations*, Oxford University Press.

Salmon, W. (1989). Four Decades of Scientific Explanation. I Kithcer P. and Salmon, W. (eds), *Scientific Explanation*. University of Minnesota Press

Schickore, J. (2011). More thoughts on HPS: Another 20 years later. Perspectives on Science, 19(4), 453–481

Schindler, S. (2013). The Kuhnian mode of HPS. *Synthese*, 190(18), 4137–4154

Schindler, S., & Scholl, R. (2020). Historical Case Studies: The "Model Organisms" of Philosophy of Science. *Erkenntnis*, 1-20.

Sorensen, R. A. (1992a). Thought experiments and the epistemology of laws. Canadian Journal of Philosophy, 22(1), 15-44.

Sorensen, R. A., 1992b, Thought Experiments, Oxford: Oxford University Press.

Sterelny, K. (2003). *Thought in a Hostile World*. Oxford University Press.

Stuart, M.T.(2016). Taming theory with Thought Experiments: Understanding and Scientific Progress", *Studies in History and Philosophy of Science*, 58: 24–33.

Stuart, M.T., Fehige, Y. and Brown, J.R. (Eds.) (2018) *The Routledge Companion to Thought Experiments*. Routledge.

Williamson, Timothy, 2004a, "Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking", Proceedings of the Aristotelian Society, CV: 1–23

Williamson, T. (2008). The Philosophy of Philosophy. Oxford University Press.

Woody, A. I. (2014). Chemistry's periodic law: Rethinking representation and explanation after the turn to practice. IN: Soler, L., Zwart, S., Lynch, M., & Israel-Jost, V. (Eds.). (2014). Science after the practice turn in the philosophy, history, and social studies of science. Routledge.

Wright, J.C. (2010). On Intuitional Stability: the Clear, the Strong, and the Paradigmatic. *Cognition*, 115(3): 491–503.

Ziolkowski, A. (2019). The Stability of Philosophical Intuitions: Failed Replications of Swain et al. (2008). Episteme, 18(2): 328-246.