Philosophical Thought Experiments as Heuristics for Theory Discovery[1]

Sara Kier Praëm & Asbjørn Steglich-Petersen

Forthcoming in *Synthese*

ABSTRACT: The growing literature on philosophical thought experiments has so far focused almost exclusively on the role of thought experiments in confirming or refuting philosophical hypotheses or theories. In this paper we draw attention to an additional and largely ignored role that thought experiments frequently play in our philosophical practice: some thought experiments do not merely serve as means for testing various philosophical hypotheses or theories, but also serve as facilitators for conceiving and articulating new ones. As we will put it, they serve as 'heuristics for theory discovery'. Our purpose in the paper is two-fold: (i) to make a case that this additional role of thought experiment deserves the attention of philosophers interested in the methodology of philosophy; (ii) to sketch a tentative taxonomy of a number of distinct ways in which philosophical thought experiments can aid (and historically have aided) theory discovery, which can guide future research on this role of thought experiments.

1. Introduction

The growing literature on philosophical thought experiments has so far focused almost exclusively on the capacity of thought experiments to help confirming or refuting philosophical hypotheses or theories. There is a lively debate on this subject, and for good reasons.[2] After all, most philosophers recognise that, as a matter of fact, they often rely on thought experiment when they attempt to amass evidence for or against philosophical theories, so it has felt urgent to many whether this reliance is in fact justified, and how exactly we should understand its nature. This sense of urgency has been amplified in recent years by suggestions and seeming evidence that thought experiments, and the intuitive judgments they

[1] Earlier versions of this paper were presented at Aarhus University (2012; 2014), University of Copenhagen (2013), and University of Southern Denmark (2013). We are grateful to the audiences on those occasions for helpful discussion, in particular Jens Christian Bjerring, Jessica Brown, Otávio Bueno, Jacob Busch, Michael Devitt, Jane Friedman, Mikkel Gerken, Raul Hakli, Brian Leiter, Hannes Leitgeb, Anna-Sara Malmgren, Stephen Mumford, Nikolaj Nottelmann, Samuel Schindler, Johanna Seibt, Asger Steffensen, Anand Vaidya, and Timothy Williamson. We are also grateful to John Hawthorne and number of referees for this journal for their useful comments on earlier versions of the manuscript. Research for this paper was funded by the Danish Council for Independent Research, as part of the project 'Epistemology of Modality: Six Investigations'. Support was also received from the John Templeton Foundation, and the 'New Insights and Directions for Religious Epistemology'-project at the University of Oxford, which one of the authors visited during Hilary and Trinity terms of 2013.

[2] See e.g. Ichikawa and Jarvis (2009), Gale (1991), Häggqvist (1996), Gendler (2004), Williamson (2007), Cohnitz (2003).

are often thought to rely on[3], may be unreliable, thus questioning a central methodological tenet of philosophy.[4] In this paper we draw attention to an additional role that some thought experiments play in our philosophical practice: some thought experiments do not merely serve as means for testing various philosophical hypotheses or theories, but also serve as facilitators for conceiving and articulating new ones.[5] As we will put it, they serve as 'heuristics for theory discovery'. That this role of thought experiments has been largely ignored in the literature is hardly surprising.[6] It is less urgent to the respectability of philosophy how theories are invented and conceived of, compared to how they might be justified. Another likely but to our mind unfounded reason why this role has been ignored, is that it may have seemed to many that we are less likely to find very systematic ways or procedures by which philosophical thought experiments aid discovery, compared to the ways in which they enable confirmation or refutation.

Our purpose in the paper is two-fold: first, to make a case that the role of thought experiments as heuristics for theory discovery deserves the attention of philosophers interested in the methodology of philosophy; and second, to sketch a tentative taxonomy of a number of distinct ways in which philosophical thought experiments can aid (and historically have aided) theory discovery, which can guide future research on this role of thought experiments. We will proceed as follows. In Section 2, we review some related claims made about the role of thought experiments in the natural sciences. In Section 3, we note a number of obvious and uninteresting ways in which philosophical thought experiments may be said to aid the discovery of new philosophical theories, and articulate a criterion for what it would take for the role of thought experiments in discovery to be methodologically interesting. In

---

[3] It is commonly held that a particular kind of (irreducible) mental state termed 'intuition' or, perhaps less often and less controversially, 'intuitive judgment', is what is doing the actual "work" in thought experimentation. A preliminary remark pertains to this. By using the term 'intuitive judgment', we wish not to commit to any theoretical possibility currently on offer -- on our intended usage it simply refers to those judgments that participants of the debate on thought experiments take themselves to be discussing. That usually means immediate responses to thought experiments. However, the focus of our paper is not so much the role of these judgments in thought experiments, but rather the role of thought experiments, more generally, in philosophical methodology.

[4] For criticism of the use of intuition in philosophy, see for instance Cummins (1998), Devitt (1994), Hintikka (1999), Kornblith (2002), (2005), (2006); and perhaps most prominently in recent years the attack from the experimental philosophy movement, see e.g. Machery, Mallon, Nichols, and Stich (2004); Stich (1998); Weinberg, Nichols, and Stich (2001); Nichols, Stich, and Weinberg (2003); and Swain, Alexander, and Weinberg (2008).

[5] It is tempting to put the distinction between these two roles of thought experiments in terms of Hans Reichenbach's famous distinction between the 'context of discovery' and the 'context of justification' (Reichenbach 1938). However, because of the complicated and controversial history of this distinction, we will stick to the more neutral description of it as two roles, and elaborate on particulars and varieties as needed.

[6] To say that this role has been ignored in the literature is not to say that theorists have been unaware of it – we make no claim to that effect. We merely note that explicit discussion of the role in the literature is so far more or less absent, and argue that the role deserves attention.

Section 4, we describe a series of thought experiments from the recent history of philosophy, and argue that they satisfy the criterion for being methodologically interesting outlined in the previous section. This serves our first main aim: to show that the role of thought experiments in discovery is sufficiently interesting to deserve the attention of philosophical methodologists. Finally, in Section 5, we sketch a taxonomy of a number of distinct ways in which philosophical thought experiments can aid theory discovery, and illustrate these with some cases from the recent history of philosophy. In particular, we distinguish between, on the one hand, a variety of ways in which thought experiments can aid the discovery of *ad hoc* repairs or developments of existing theories, and, on the other hand, a more open-ended potential for theory discovery exhibited by other kinds of thought experiments.

2. Parallel claims about scientific thought experiments

While the use of thought experiments in aid of discovery has been overlooked in the debate on *philosophical* methodology, the idea has been applied to thought experiments as employed in the natural sciences. Brown (1986; 1991a; 1991b) proposes a taxonomy of thought experiments in physics, according to which they can be either 'destructive' or 'constructive'; the former kind 'destroys or at least presents serious problems for a theory' (Brown 1991a: 34), while the latter kind aids understanding, development, or initiation of theories (Brown 1991a: 36-43). Brown claims that a small number of thought experiments, so-called 'platonic' thought experiments, are both destructive and constructive, and so can, at the same time, destroy an old theory and aid the creation of a new one. According to Brown, such thought experiments provide a priori access to knowledge about laws of nature, where these are understood as independently existing abstract objects in the form of relations between universals. While Brown is the most prominent proponent of the idea that thought experiments can aid discovery, he is not the only one. A number of papers by Norton both challenges and complements Brown's view (1991; 2002; 2004): challenges it in that Norton denies Brown's claim that thought experiments provide direct access to a Platonic world ('thought experiments in science are merely picturesque argumentation', Norton says (2004: 1139)); complements it in that he too, despite advocating a deflationary empiricist view on thought experiments, believes thought experiments to be a reliable mode of inquiry for generating new knowledge. In fact, Norton explicitly sees thought experimentation as serving both justification and discovery, although the way in which he views the justification process is quite different from Brown's understanding. On one reading, Norton could be said to reject the view that thought experiments provide justification; more precisely, he holds that it is the

reconstruction as an argument that functions as the justification. The thought experimental outcome is 'justified only insofar as the reconstructed argument can justify the conclusion'. However, to the extent thought experiments are considered (disguised) arguments, then thought experiments provide justification.

Nothing in our approach hinges on identifying thought experiments as either platonic or as arguments. Our argumentative emphasis and the domain of interest in this paper are different from that of Brown and Norton; they both confine their claims to thought experiments as employed in the sciences, and whether or not either account could be extended to encompass *philosophical* thought experiments is unclear. As we understand philosophical thought experiments, they concern metaphysical possibilities, whereas thought experiments in the natural sciences operate by uncovering physical possibilities, such that, for example, an unsupported object will fall.[7] We shall focus on the use of thought experiments in philosophy exclusively. Only in philosophy do we find such a methodologically significant widespread use of thought experiments and it is this methodology that we are primarily interested in. While we acknowledge Brown and Norton's contributions to understanding the creative role of thought experiments in the sciences,[8] we therefore leave behind their approaches for now in order to focus on the peculiarities of philosophical thought experiments (henceforth simply 'thought experiments').

3. When is the role of thought experiments in discovery methodologically interesting?

That thought experiments can play a role in the discovery of new hypotheses or theories is in itself unremarkable. Virtually anything can, under the right circumstances, aid discovery – a hearty breakfast, a train of thought triggered by an overheard conversation, a bump on the head, etc., are all events that could result in someone conceiving of a new theory, without this being in any way methodologically interesting. If these sorts of events can aid discovery, so can contemplating a thought experiment, but that would not in itself be of particular interest. So anyone wishing to argue that the creative role of thought experiments in conceiving of new theories deserves more attention has to provide a criterion that allows one to distinguish the methodologically interesting cases from the uninteresting ones. That will be the purpose of

[7] In part inspired by Bealer (1998), who contrasts *rational* and *physical* intuitions as elicited by thought experimentation in philosophy and natural sciences, respectively (Although Bealer thinks the term 'thought experiment' should be abandoned in philosophy).

[8] An alternative take on how to understand the discovery-functioning of scientific thought experiments is put forward by Kuhn. According to Kuhn (1964), scientific thought experimentation serves as a discovery-tool in that it produces new understanding *via* a reconceptualization of old empirical data. It is unclear, however, what the relevance of this approach would be for understanding the creative potential of *philosophical* thought experiments.

this section.

We can approach this by considering a variety of methodologically uninteresting cases that we want our criterion to rule out. Consider first the anecdote of how Newton conceived of his theory of gravity by having an apple fall on his head, causing a stroke of brilliant insight. Although the anecdote is almost certain to be false or at least wildly exaggerated, it is in principle conceivable that such an unusual causal process could have taken place. Similar unusual processes also easily come to mind, and we suspect that many people who are engaged in research have experienced how some event unconnected to one's topic of research can cause one to have thoughts that end up advancing one's theorising. Although striking, such events are not of methodological interest, and we thus want our criterion to rule them out. One possible way of doing so would be by demanding that in the interesting cases, what aids discovery must do so with a certain degree of regularity, not in the sense of by itself tending to be a sufficient cause of discovery (nothing is likely to play such a strong role), but in the sense of regularly playing a causal role in some larger set of causes of discovery. That would certainly rule out the story of Newton's discovery as methodologically interesting; if having apples fall on one's head regularly aided brilliant insight, science would be far more advanced than it is!

Someone might doubt that this should be a necessary condition: why think that the role of thought experiments in discovery couldn't be methodologically interesting, even if it were quite rare that they aided discovery in the relevant way? As it happens, and as we will go on to support, we think it is a very common occurrence that thought experiments play this role. But we also think that it should be part of the criterion for being methodologically interesting, since a methodology should supply methods or procedures that one could rely on to aid discovery with some expectation of regularity. It should be immediately clear, however, that this kind of regularity by itself does not suffice as a criterion. It seems plausible, for instance, that having a hearty breakfast (or drinking coffee, or getting a good night's sleep, or what have you) before engaging with one's research regularly aids discovery without thereby being of methodological interest. One way of ruling out such cases would be to demand that methodologically interesting aids to discovery not only regularly play a role in causing discovery, but does so in a way that enables prediction or anticipation, or at the very least makes it unsurprising, *what* new theory or hypothesis is discovered. That would rule out hearty breakfasts, etc., as methodologically interesting aids of discovery. Although they regularly aid discovery, they do not, when considering the set of causes of some discovery, make it unsurprising *what* specific discovery was made, but only help make it unsurprising

that *some* discovery, whatever its content, was made. This criterion also rules out as interesting a further trivial way in which thought experiments might aid discovery, namely aid in the purely negative sense of refuting old and mistaken theories, thus clearing the way for new ones.

This, then, shall be our guiding criterion in the following: that the consideration of a thought experiment aided the discovery of a new theory or hypothesis is methodologically interesting to the extent that considering the thought experiment made it unsurprising what theory or hypothesis was discovered, and to the extend that this kind of aid is instantiated with some regularity.

It should be acknowledged immediately that this criterion is quite vague – what does it take, for example, for it to be 'unsurprising' what theory or hypothesis was conceived of as a result of considering a thought experiment? Part of this will become clearer when we go on to consider examples of cases that we think satisfy the criterion. But it should also be pointed out that some degree of vagueness in the criterion is likely to be unavoidable, since the phenomenon it targets is not itself very precise. One virtue of the criterion is that it allows for degrees of methodological interest, so the role of thought experiments in discovery can vary in interest from case to case.

We should also point out some features that our criterion does *not* require for methodological interest. It does not require that considering the thought experiment was *essential* to the relevant discovery, in the sense that the discovery couldn't have been made in the absence of the experiment. Nor does it require that the thought experiment was originally constructed with the purpose of aiding the discovery of a new theory or hypothesis. It is fully compatible with our criterion, for example, that a thought example aids discovery in a methodologically interesting way, even if it was originally constructed with another purpose, such as that of refuting another theory. The criterion also doesn't require that consideration of thought experiments be a fool proof or perfectly reliable aid to theory discovery, as long as it aids with some degree of regularity. These limitations collectively motivate our preferred terminology of thought experiments as a 'heuristic' for theory discovery: an inessential, sometimes accidental, and fallible, but nevertheless useful method for discovering new philosophical theories or hypotheses.

With this criterion in place, we can move on to considering whether thought experiments ever satisfy it, thus making their role in discovery methodologically interesting. If such cases can be found, that would be a good argument for devoting more interest to the role of thought experiments in discovery, than it currently receives.

4. Examples of thought experiments acting as a discovery heuristics

As our example of a thought experiment that plays a role for discovery in the relevant sense, we have chosen Gettier's classic counterexample to the traditional analysis of knowledge as justified true belief, and some of the thought experiments that Gettier's original examples have provoked (Gettier 1963). We have chosen these as our central case in this section for three main reasons. The first reason is that the importance and influence of the Gettier cases in a central field of philosophy stands undisputed. They are, as many have noted, paradigmatic cases of successful thought experiments. The second reason is that every commentator that we are aware of focus on the role of these thought experiments in testing the validity of philosophical theories only. As mentioned above, we think that the very same thought experiment can at the same time serve to test theories, and play a role in theory discovery that is interesting on our criterion. Considering the Gettier cases brings this feature out. The third reason is that the long series of new analyses followed by new counterexamples exhibited by the Gettier literature nicely illustrates the role of thought experiments in developing theory. While we don't think that the type of discovery provoked by Gettier cases is the most interesting kind of discovery that thought experiments can aid (we will describe some more interesting ones later on), we therefore do think that these cases are especially well suited to the purpose of this section, i.e. to show that there are cases of theory discovery aided by thought experiments in a methodologically interesting way.

In his influential contribution, Gettier questions the traditional account of the necessary and sufficient conditions for someone's knowing a given proposition P: S knows that P, if and only if (i) P is true, (ii) S believes that P, and (iii) S is justified in believing that P. To refute this analysis, Gettier constructs two thought experiments engendered to show that this analysis is insufficient. Since the experiments are similar, we shall focus on one only.

We are asked to consider a scenario in which Smith and Jones both have applied for the same job. In this scenario, we are provided with details that make us judge that Smith has a justified belief that Jones will get the job. Smith also believes that Jones has ten coins in his pocket, which is true and demonstrated to him, and so Smith has strong evidence for the following conjunctive proposition: (d) Jones is the man who will get the job, and Jones has ten coins in his pocket. From this Smith infers, that (e) the man who will get the job has ten coins in his pocket, and accepts (e) on the basis of (d) for which he has strong evidence. Clearly, Smith is justified in believing that (e) is true. As it turns out, Smith is wrong that Jones will get the job; Smith is offered the job and not Jones. In addition, unbeknownst to Smith, he, too, has ten coins in his pocket. So (e) is true even though the proposition from

which it was inferred, (d), is false. In this example, then, all of following is true: (i) (e) is true; (ii) Smith believes that (e) is true; and (iii) Smith is justified in believing that (e) is true. But, Gettier claims, it is clear that Smith does not *know* that (e) is true, because his belief is based on the number of coins in Jones' pocket, whom he falsely believes is the man to get the job, while it is, in fact, in virtue of the number of coins in Smith's own pocket (a number entirely unbeknownst to Smith) that (e) is true. Gettier concludes that this thought experiment shows that the traditional analysis fails to state sufficient conditions for someone's knowing a given proposition.

This description only makes apparent the role Gettier himself took his experiments to play, namely that of testing – and refuting – the traditional analysis. To see the potential of Gettier's cases for the discovery of new theories, we must look at the many responses it triggered. Since the publication of Gettier's cases, a large number of successive analyses of knowledge have been presented to avoid Gettier's original counterexamples, as well as new counterexamples targeting further analyses. By focusing on a few cycles of counterexamples and new analyses, it becomes apparent that the thought experiments intended as counterexamples also play a role in the discovery of new analyses of knowledge that satisfy our criterion of methodological interest, outlined in the previous section.

Consider the first new analysis proposed in response to the Gettier cases. When reading through Gettier's counterexample, one not only realises that the original analysis fails. The counterexample also makes it obvious *why* it fails, at least in this particular case – indeed Gettier himself points this out: Smith infers his true belief from a false one, and it seems clear enough that this is the reason we are reluctant to assign knowledge in this particular case. Having made this diagnosis, however, an obvious solution suggests itself (we will discuss exactly how in the next section): amend the analysis of knowledge to rule out the feature which, *in this particular case*, stands in the way of knowledge, by restricting attribution of knowledge to instances of justified true belief where the justification for the belief does not rest on any false grounds. And indeed, this 'No False Grounds' analysis (NFG) was the first amended analysis of knowledge proposed in response to Gettier (Clark 1963).

A variety of counterexamples emerged, however, showing that adding a No False Grounds clause still does not result in a satisfactory definition of knowledge.[9] Take, for instance, Lehrer's (1965) example: Suppose I have a justified false belief that Mr. Nogot, who

[9] In fact, Clark (1963) himself shows by method of counterexample that 'Justified True Belief + No False Ground' is not an adequate analysis of knowledge, before advancing his own analysis ('Justified True Belief + No *Essential* False Ground') (see below).

is in my office, owns a Ford. I have an equally justified true belief that Mr. Havit, also in my office, owns a Ford. From both of these beliefs I infer the justified true belief that *someone* in my office owns a Ford. But since this belief is partly based on the false belief that Mr. Nogot owns a Ford, the No False Grounds analysis gives the result that I fail to *know* that someone in my office owns a Ford. But this seems wrong. As long as I have at least one independent series of all true justification for P, I know, regardless of any other false evidence for P I might possess. Hence, the No False Grounds analysis fails as an adequate analysis of knowledge on account of being too strong: it is possible for a person to know that P without satisfying all the conditions stated in the analysis as necessary, that is, it entails that a person fails to know a proposition whenever any part—even a dispensable part—of the justification is false.

But again, from considering Lehrer's example, we seem to learn more than the failure of the No False Grounds analysis. The example also makes it obvious *why* it fails in this particular case, thus suggesting a simple repair. The No False Grounds analysis requires us to deny a person knowledge when this person has two independent series of justification; a justified-but-false series and a justified-but-true series. But it seems clear from the thought experiment that in this particular case, the person should be able to omit the former of these series of justification and rely exclusively on the latter with the result of still being justified in believing P and so have knowledge that P. And so, by way of this thought experiment, this particular case constructed by Lehrer makes us aware of how the No False Grounds analysis might be corrected to deal with this particular case; we ought to replace the supplementary condition of No False Grounds with a so-called 'No *Essential* False Grounds' clause—that is, knowledge is a justified true belief, where the justification for the belief does not *essentially* rest on any false beliefs.'

This analysis too was quickly subjected to a variety of counter examples.[10] But where the No False Grounds analysis was charged with being too strong, the No Essential False Grounds analysis was charged with being too weak. Consider the following thought experiment, due to Skyrms (1967): Suppose a pyromaniac has just purchased a box of Sure-Fire Matches. He has observed that Sure-Fire matches always light when struck. The

[10] In fact, both the 'Justified True Belief + No False Ground' analysis and 'Justified True Belief + No *Essential* False Ground' analysis generated a plethora of responses. Other counterexamples, some of which intended to target the sufficiency of not only one, but both additions, include: Lehrer's 'Non-inferential Nogot' (1965), (1970); Lehrer's 'Clever Reasoner' (1974); Feldman's 'Testimony Nogot' (1974); Scheffler's 'Stopped Clock' (1965) following Russell (1948); Chisholm's 'Sheep in the Field' (1966); Skyrms' 'Sure Fire Match' (1967) and Rozeboom's 'Togethersmith' (1967). Saunders and Champawat (1964) together with Skyrms (1967) question the necessity of the additions.

pyromaniac is justified in believing that the match he is holding will light when struck. It does. And so he has a justified, true belief. Unbeknownst to the pyromaniac, however, his match was defective and had an incredibly rare impurity that prevented the match from being lit by friction, but it lit anyway due to a sudden burst of Q-radiation igniting the match. Although the conditions stated in the No Essential False Grounds analysis are fulfilled, again we are reluctant to assign to the pyromaniac knowledge. Although he has a justified true belief that does not depend essentially on any false beliefs, his belief is true only accidentially. Hence, the analysis fails on account of being too weak: it is possible for a person to fail to know that P even when satisfying all the conditions stated in the analysis as sufficient; a person can fail to know that P even when she has a justified true belief that P, based entirely on known, true evidence.

Again, we learn more than the failure of Lehrer's analysis by considering this case. We learn *why* it fails in this particular instance: the truth of the pyromaniac's justified belief that the match will light only accidentally obtains as a result of a sudden burst of Q-radiation. Making the obvious generalisation, the thought experiment[11] features a true proposition, a so-called *defeater*, unbeknownst to the pyromaniac, such that were that proposition to be added to the rest of his evidence, he would no longer be justified in the relevant belief. If only the fact that the match was defective had been known to the pyromaniac, he would no longer have justification for believing that the match would light. And so, by way of this thought experiment, this particular case constructed by Skyrms makes us aware of how Lehrer's analysis might be corrected; we ought to replace the No Essential False Grounds condition with a so-called 'No Defeaters' condition—that is, knowledge is a justified true belief, where there are no defeaters of S's justification for P.

And so yet another candidate for an analysis of knowledge is born. The No Defeaters approach, too, was opposed with a cascade of counterexamples. In fact, the majority of the attempts of solve the Gettier problem developed as reactions to particular counterexamples, only to become target themselves of further counterexamples. All told, Gettier's examples initiated a period of enormous epistemological innovation in attempt to repair or replace the traditional definition of knowledge. The here mentioned proposals represent just a few; other notable proposals include various Externalist approaches, Sensitivity strategies and Safety strategies. While advocates of the various theoretical innovations following Gettier might still

[11] And similarly constructed thought experiments; Gettier's original cases would work just as well. Our project here is not to claim that Skyrms' thought experiment is the single origin of the various theories of defeaters. In fact, we are sure it is not.

endorse them, the general consensus is that none of the attempts so far have succeeded in solving the Gettier problem. Indeed, even the seeming intractability of the Gettier problem has been subject to investigation (see e.g. Zagzebski (1994) and Craig (1990)).

The purpose of this section was to identify examples of thought experiments playing a role in discovery, satisfying the criterion for methodological interest sketched in Section 3. We submit that the above three thought experiments played such a role, at least in the ways exemplified above. Gettier's original cases played an important role in the discovery of the No False Grounds analysis, as did Lehrer's case in the discovery of the No Essential False Grounds analysis, and Skyrm's case in the discovery of the No Defeater analysis. Considering the details of each thought experiment and analysis, it is entirely unsurprising that those particular analyses were discovered in response to those particular thought experiments. Furthermore, the general pattern of counterexamples acting as cradles for new theories that the three examples are part of shows that thought experiments act in this role with the regularity needed for methodological interest. If this is plausible, it by and large fulfils the first general aim of this article, namely to argue that the role of thought experiments as heuristics in the discovery of new hypotheses and theories is sufficiently interesting to warrant more interest than it currently receives.

Before moving on, there is an important objection to consider. When considering the sequence of thought experiments and discoveries above, someone might object that the order of events is actually reversed: in reality, first comes the theory and then comes the thought experiment. So, it might be claimed, it is not that a particular thought experiment aids discovery of a new theory, but rather that the conception of a new theory prompts the construction of a particularly well suited thought experiment, engineered so that it not only works as a counterexample to an already existing theory, but at the very same time lends support to the new but already conceived theory. Perhaps the No Defeaters analysis was not discovered by considering the counterexample to the No Essential False Grounds analysis, but rather thought of prior to detecting the counterexample. Perhaps the No Essential False Grounds analysis was not discovered by considering the counterexample to the No False Grounds analysis, but rather thought of prior to detecting the counterexample. And so on.

We admit that it is possible that this alternative reconstruction of events is accurate; it is certainly true that philosophers sometimes construct new theories (with or without the aid of thought experiments) and then engineer thought experiments to support the theory and at the same time refute others. It may be difficult to distinguish subsequently which thought experiments have been constructed to serve this latter *post hoc* function, and which thought

experiments aid discovery, but not altogether impossible. Perhaps by studying more instances of thought experiments seemingly falling in either of these classes, we will learn to discern some general characteristics and so understand how to separate the two categories more easily. But in any case, the difficulty of discerning these two types does not imply that the cases we are interested in never occur. And we think it is hard to deny that reflection on cases that were intended purely as counterexamples at least sometimes (not to say often) aid the conception of new theories in the way described.

5. A taxonomy of thought experiments as heuristics for theory discovery

In this last main section, we sketch a tentative catalogue of a number of distinct ways in which philosophical thought experiments can aid theory discovery, and illustrate these with some cases from the recent history of philosophy. In particular, we distinguish between, on the one hand, a variety of ways in which thought experiments can aid the discovery of *ad hoc* repairs or developments of existing theories, and, on the other hand, a more open-ended potential for theory discovery sometimes exhibited by other types of thought experiments. We will also return to the pressing question, postponed from Section 4, of *how* exactly thought experiments can help one conceiving of new or improved theories; so far, we have merely argued *that* they sometimes play this role. It should be emphasised that what we provide in this section is not a catalogue of different kinds of *thought experiments*, but of some distinct *ways* in which thought experiments can aid discovery. What we say is therefore open to the possibility that some thought experiments can aid discovery in several of the ways to be sketched at once. Our observations here will be quite tentative; we hope, however, that they will serve as a starting point and guide for future research in this important but overlooked area.

5.1 Counterexamples as heuristics for discovery

As mentioned, the first and simplest way in which philosophical thought experiments can aid theory discovery, is the manner in which thought experiments that are intended mainly as counterexamples to some theory can act as cradles for improved theories, exemplified in the previous section. Since we have already described several such cases in some detail, we shall focus on those cases in this section as well. There are two main kinds to consider: counterexamples to the sufficiency of some philosophical theory or analysis, and counterexamples to necessity. Below, we provide some initial proposals of how to understand the characteristic ways in which these kinds of thought experiments aid discovery.

5.1.1 Counterexamples to sufficiency

How do counterexamples to the *sufficiency* of some theory or analysis, such as those presented by Gettier, sometimes make it apparent how the theory or analysis ought to be improved, or with what they ought to be replaced? When considering Gettier's own example about Smith and Jones competing for a job, we noted that it seems obvious that what's hindering knowledge in this particular case is the fact the knowledge-aspiring belief is based on a falsehood, which in turn suggests the obvious repair, namely to amend the analysis of knowledge so as to rule out beliefs based on falsehoods. A promising explanation of the ease with which the potential solution suggests itself relies on the fact that it is much easier to see whether some property obtains in concrete cases, and to see what concrete factors hinder the property from obtaining if it doesn't, than to state general conditions for the property to obtain. Counterexamples to sufficiency tend to provide concrete cases where the presence of some concrete factor hinders the conditions of the target analysis in resulting in the target property. In many cases, it will be a relatively trivial matter to discern the concrete hindering factor. Not as trivial as discerning factors directly mentioned by the case, such as the names of the protagonists Smith and Jones, or the fact that they are both applying for the same job. Some modicum of inference is required to discern the hindering factor. In the case at hand, the inference might involve imaginatively deleting the hindering factor from the case, to see if that makes a difference to our judgment of knowledge. What is *not* trivial, and what thought experiments of this sort are less helpful with, is to discern the level of generality at which the amended analysis should rule out the hindering factor. In the case under consideration, amending the analysis by directly banning beliefs based on falsehoods yields the correct answer. But as further thought experiments showed, this amendment was not at the appropriate level of generality to rule out further counterexamples. As they show, reliance on a falsehood only prevents beliefs from qualifying as knowledge in cases where the reliance is an essential part of the belief's justification. This reveals the necessity of repeated experiment in refining a theory or approach, but it nevertheless remains true that the thought experiment helped us improve our analysis by providing a concrete case highlighting a concrete hindering factor to be ruled out by the improved analysis.

Not all counterexamples to sufficiency make clear how to improve the analysis. This will be the case, for example, when the counterexample speaks against a particularly fundamental presupposition of a theory, rather than less fundamental tenet of it. A striking example is Frank Jackson's (1986) case of Mary, the brilliant neurophysiologist, who despite knowing everything there is to know about what goes on physically when we experience

colours, still discovers something new when she is released from the black and white room in which she has been kept since birth, and experiences colours herself for the first time. The thought example seemingly demonstrates the shortcoming of physicalist theories of conscious experience. Since Mary knew everything to be known about the physical side of conscious experience, but still discovered something new, there is more to know about conscious experience than contained by the physical information. But the thought experiment doesn't make obvious any quick recovery for the physicalist theory, and therefore doesn't help us improve our theory of consciousness, beyond the merely negative help of ruling out physicalism as the correct theory.

5.1.2 Counterexamples to necessity

Consider now thought experiments that function as counterexamples to the *necessity* of some theory or analysis, such as Lehrer's counterexample to the No False Grounds analysis described above. We think that on a general level, such thought experiments aid theory improvement in much the same manner as counterexamples to sufficiency. In Lehrer's case, the target analysis fails not because of the presence of a concrete factor hindering knowledge as in Gettier's own example, but because of a concrete factor cancelling out the relevance of a factor deemed to prevent knowledge by the No False Grounds analysis, namely a falsehood among the justifying beliefs. As a result, we need to amend or replace our analysis with one that does without or replaces whatever condition is shown to be irrelevant. As noted above, in Lehrer's case, one possible repair suggests itself, namely to restrict the ban on false grounds to those that play an essential or indispensable justifying role. The counterexample aids the discovery of this improvement by providing a concrete case with a concrete factor that renders a condition irrelevant, and in such concrete cases, it is a relatively easy task to introduce amendments that, in that particular case, renders the analysis correct. Again, the method by no means guarantees the right result. It is hard to know from such concrete cases whether the improvement is pitched at the right level of generality, thus making further thought examples necessary. But again, it remains true that the thought experiment helped us improve our analysis by providing a concrete case highlighting a concrete hindering factor, that it would have been difficult to anticipate the relevance of by mere abstract reflection on the target concept.

As above, there are also cases of counterexamples to the necessity of analyses that do not in the same way make it obvious how the relevant analysis should be amended or replaced. Striking examples include cases of pre-emptive causation, targeting various

philosophical accounts of causation that appear incompatible with them. Simple counterfactual accounts of causation make it a necessary condition for causal relationships that the effect counterfactually depends on the cause, such that had the cause-event not occurred, neither would the effect-event. But a series of thought experiments has demonstrated that this is not necessary. The actual cause might pre-empt another cause that would have caused the effect to come about even if the actual cause-event had not occurred. As was the case with Jackson's case of Mary, these counterexamples seem indicative of a more fundamental flaw than can be fixed by adding constraints on the relevant counterfactual relation, thus making the counterexamples less useful as heuristics for theory discovery (for an overview of this debate, see Hall and Paul 2013).

5.2 Open-ended thought experiments

While counterexamples can act as heuristics for discovering *ad hoc* fixes to existing theories, other thought experiments seem to aid theory discovery in more fundamental and open-ended ways. This role of thought experiments is likely to be much more heterogeneous, but we can bring it into focus by comparing it to the role thought experiments can play in adding *confirmation* to a theory – the role played by what is sometimes called 'constructive' thought experiments. Consider Thomson's famous case of the violinist (1971), singled out by Brown and Fehige (2014) as a thought experiment of that constructive kind. Thomson asks you to imagine finding yourself waking up one morning in bed with an unconscious famous violinist who has had his circulatory system plugged into yours, so that your kidneys can extract poison from his blood as well as your own. If he is unplugged, he will die, but in nine months he will have recovered and can be safely unplugged. The point of the example is to support the thesis that women have a right to abortion, even granting that foetuses have a right to life. Most of us would be inclined to judge that even though the famous violinist has a right to life, you would have a right to deny him the use of your body as his life support system, and it seems plausible that, at least in some cases, women's relation to the foetus in their wombs are relevantly similar. Another striking example is Rawls' thought experiment (and similar thought experiments from the social contract tradition in political philosophy) asking us to imagine being placed behind a 'veil of ignorance', where one does not know one's place in society or one's fortune in the distribution of natural assets and abilities (Rawls 1999: 118). From this point to view, Rawls conjectures that we would be inclined to accept certain egalitarian principles of justice, and that these judgments would carry special weight because of the unbiased standpoint from which they are made. Yet another example is Putnam's

(1973) famous Twin Earth thought experiment, designed to support semantic externalism. Putnam asks us to consider a planet exactly like Earth elsewhere in the universe, with twin equivalents of every person and thing on Earth, the only difference being that there is no water on Twin Earth, but instead a liquid that is superficially identical, but chemically different, in not being composed of H20. Putnam thinks that given this, we would be inclined to judge that when earthlings and their Twin Earth counterparts say 'water', they mean different things, even though there is no internal difference to be found between the earthlings and their counterparts, thus leading Putnam to conclude that meanings are determined in part by factors external to language-users.

In all of these cases of constructive thought experiments, the theories they are taken to support seem, in the main, to have been conceived of or discovered *prior* to the construction of the experiment. The thought experiments may have played a role in *convincing* Thomson of the moral permissibility of abortion, Rawls of his proposed principles of justice, and Putnam of semantic externalism, but it is hard to imagine that they constructed the experiments in a spirit of genuine open-mindedness about the relevant views with the purpose of making new discoveries, or that they were surprised by what the thought experiments revealed once constructed. Others may be surprised upon considering their thought experiments; they may be surprised, for example, that a thesis they had found implausible receives support by the experiment. But that just shows that thought experiments lending support to theories have one of their main uses in arguments designed to convince others of the theories, and perhaps add to their understanding of these theories by providing illustrations of them.

With this comparison with 'constructive' thought experiments in place, the open-ended role of thought experiments in discovery can be brought into focus. What we are after are cases where we construct or consider a thought experiment, not (just) in order to add support or details to a theory or hypothesis that one has already in the main conceived of, but in a more open-minded spirit of seeking answers to questions that one is genuinely in doubt about. We think that such cases can be found mainly in the form of modifications of previous thought experiments, using those experiments as templates for new ones. The many variations of Putnam's Twin Earth are examples of this. Impressed by the potential of imagining a Twin Earth in establishing semantic externalism, other authors have probed the Twin Earth scenario to ask new questions: Do *mental states* depend on external factors (McGinn 1977)? Are *moral properties* natural kinds (Horgan and Timmons 1992)? Does the *epistemic status* of our beliefs depend on external features of a kind revealed by Twin Earth scenarios (e.g.

Williamson 2000)? Etc. What these examples illustrate is the potential of a certain scenario – a planet identical to ours in all but a few respects – in uncovering dependencies or the lack thereof between a number of psychological, semantic, moral, and epistemic properties, and properties that are external to the agents implicated in these properties. In these cases, it seems plausible that the general template of the Twin Earth thought experiment played an important role in the discovery of theories that it was not originally intended as evidence for. That is, even if Putnam came up with the Twin Earth case primarily with the aim of lending illustration and evidential support to a theory that he had already conceived of, the template of the thought experiment subsequently played important roles in the discovery other philosophers made of parallel theories in other areas of philosophy. We cannot here explore the precise nature of this role, but hope to have made it plausible that thought experiments can act as heuristics for theory discovery in this more open-ended sense.

6. Concluding remarks

We have argued that the role of thought experiments as heuristics for theory discovery deserves the attention of philosophers interested in the methodology of philosophy. This role of thought experiments is methodologically interesting in a way that the role of a hearty breakfast or a good night's sleep is not, because consideration of thought experiments regularly make it unsurprising what theory or hypothesis is subsequently discovered, and because there are systematic and distinct general ways in which philosophical thought experiments can aid (and historically have aided) theory discovery, some of which we have outlined in the previous section. In this paper we have, by necessity, only been able to take a modest step towards a more substantive understanding of the role of philosophical thought experiments in discovery, but we hope to have encouraged further research on this undeservedly neglected topic.

Bibliography

Bealer, G. (1998), "Intuition and the Autonomy of Philosophy", in Rethinking Intuition, eds. DePaul and Ramsey, Rowman &Littlefiled Publishers, pp. 201-239.

Brown, James R. (1986), "Thought Experiments since the Scientific Revolution", *International Studies in the Philosophy of Science*, 1, pp. 1–15.

Brown, James R. (1991a), *Laboratory of the Mind: Thought Experiments in the Natural Sciences*. (London: Routledge)

Brown, James R. (1991b), “Thought Experiments: A Platonic Account” in: T. Horowitz and G. Massey (eds.), *Thought Experiments in Science and Philosophy*, (Lanham: Rowman & Littlefield), pp. 119–128

Brown, James R. and Fehige, Yiftach (2014), ‘Thought experiments’, *The Stanford Encyclopedia of Philosophy*.

Chisholm, R.M. (1966), *Theory of Knowledge*. (Englewood Cliffs: Prentice-Hall).

Clark, M. (1963). ‘Knowledge and Grounds: A Comment on Mr. Gettier's Paper’, *Analysis*, 24, pp. 46-48.

Cohnitz, Daniel (2003), “Modal Skepticism: Philosophical Thought Experiments and Modal Epistemology.” In: *The Vienna Circle and Logical Positivism* [Vienna Circle Institute Yearbook 10/2002]. (Dordrecht: Kluwer Academic Publishers), pp. 281-296.

Craig, E. (1990). Knowledge and the State of Nature. Oxford: Clarendon Press.

Cummins, Robert (1998): ‘Reflection on Reflective Equilibrium’, in DePaul, Michael and Ramsey, William (eds.) (1998): *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*, Rowman & Littlefield.

Devitt, Michael (1994): ‘The Methodology of Naturalistic Semantics’, *Journal of Philosophy*, 91, 545 – 572.

Feldman, R. (1974). ‘An Alleged Defect in Gettier Counter-Examples’. *Australasian Journal of Philosophy*, 52, pp. 68-69

Gale, Richard M. (1991), ”On Some Pernicious Thought-Experiments”, in Horowitz, T. & Massey, G. (eds.), *Thought Experiments in Science and Philosophy*, (Lanham, MD: Rowman &Littlefield).

Gettier, Edmund. L. 1963: “Is Justified True Belief Knowledge?", Analysis 23, 121-123.

Gendler, Tamar (2004), ”Though Experiments rethought – and repercieved”, *Philosophy of Science* 71, pp. 1152-1163.

Hall, N., and Paul, L.A. (2013), *Causation: A User’s Guide*, Oxford University Press.

Hintikka, Jaakko (1999): ‘The Emperor’s New Intuitions’, *Journal of Philosophy*, 96, pp. 127 –147.

Horgan, T. and Timmons, M. (1992), ‘Troubles on Moral Twin Earth: Moral Queerness Revived’, *Synthese*, 92, pp. 221-60.

Häggqvist, Sören (1996), *Thought Experiments in Philosophy,* (Stockholm: Almqvist and Wiksell).

Ichikawa, Jonathan & Jarvis, Benjamin (2009), “Thought-experiment intuitions and truth in fiction”, *Philosophical Studies* 142 (2), pp. 221 – 246.

Jackson, F. (1986), 'What Mary Didn't Know', *Journal of Philosophy*, 83, pp. 291-95.

Kornblith, Hilary (2002): *Knowledge and its Place in Nature*, Oxford University Press, 2002.

Kornblith, Hilary (2005): 'Replies to Alvin Goldman, Martin Kusch and William Talbott,' in *Philosophy and Phenomenological Research*, 71, pp. 427 – 441.

Kornblith, Hilary (2006): 'Appeals to Intuition and the Ambitions of Epistemology', in Heatherington, Stephen (ed.), *Epistemology Futures*, Oxford: Oxford University Press, pp. 10 –25.

Kuhn, Thomas (1964): 'A Function for Thought Experiments', in Kuhn (1977), *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Chicago: University of Chicago Press, pp. 242-265.

Lehrer, Keith (1965), 'Knowledge,Truth and Evidence'. *Analysis*, 25, pp.168-75.

Lehrer, Keith (1970), 'The Fourth Condition of Knowledge: A Defense'. *Review of Metaphysics*, 24, pp. 122-8.

Lehrer, Keith (1974), *Knowledge*. Oxford: Clarendon Press.

Machery, Edouard; Mallon, Ron; Nichols, Shaun & Stich, Stephen (2004): 'Semantics, Cross-Cultural Style', *Cognition*, 92, B1 – B12.

McGinn, C. (1977), 'Charity, Interpretation, and Belief', *Journal of Philosophy*, 74, pp. 521-35.

Nichols, Shaun; Stich, Stephen & Weinberg, Jonathan (2003): 'Metaskepticism: Meditations in Ethno-Epistemology', in Luper, Steven (ed.) *The Skeptics*, Burlington, VT: Ashgate, pp. 227 –247.

Norton, John (1991), "Thought Experiments in Einstein's work" in: T. Horowitz and G. Massey (eds.), *Thought Experiments in Science and Philosophy*, (Lanham: Rowman & Littlefield).

Norton, John (2002), "Why thought experiments do not transcend empiricism", in C. Hitchcock, ed., *Contemporary Debates in the Philosophy of Science*. Oxford: Blackwell, pp. 44-66.

Norton, John (2004), "On Thought Experiments: Is There More to the Argument?" *Proceedings of the 2002 Biennial Meeting of the Philosophy of Science Association, Philosophy of Science,*71, pp. 1139-1151.

Putnam, H. (1973), 'Meaning and Reference', *Journal of Philosophy*, 70, pp. 699-711.

Rawls, J. (1999), *A Theory of Justice*, Belknap Press.

Reichenbach, Hans (1938), *Experience and Prediction. Analysis of the Foundations and Structure of Knowledge* (Chicago: University of Chicago Press).

Rozeboom, W.W. (1967), 'Why I Know So Much More than You Do'. *American Philosophical Quarterly*, 4, pp. 281-90.

Russell, Bertrand (1948), *Human Knowledge: Its Scope and Limits*. (New York: Allen and Unwin).

Saunders, J.T., and Champawat, N. (1964), 'Mr. Clark's Definition of "Knowledge"'. *Analysis*, 25, pp. 8-9.

Scheffler, Israel (1965), Conditions of Knowledge. (Chicago: Scott, Foresman).

Skyrms, B. (1967), 'The Explication of "X knows that p"'. *Journal of Philosophy*, 64, pp. 373-89.

Stich, Stephen (1998): 'Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity', *Synthese*, 74, pp. 391 – 413.

Swain, Stacey; Alexander, Joshua & Weinberg, Jonathan (2008): 'The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp', *Philosophy and Phenomenological Research*,76, pp. 138 – 155.

Thomson, J. (1971), 'A Defence of Abortion', *Philosophy and Public Affairs* 1, pp. 47-66.

Williamson, Timothy (2000), *Knowledge and its Limits*, Oxford University Press.

Williamson, Timothy (2007), *The Philosophy of Philosophy*, (Oxford: Blackwell).

Weinberg, Jonathan; Nichols Shaun & Stich, Stephen (2001): 'Normativity and Epistemic Intuitions', in *Philosophical Topics*, 29, pp. 429 – 460.

Zagzebski, L. (1994), 'The Inescapability of Gettier Problems,' *Philosophical Quarterly* 44, pp. 65-73.