Axiomathes (2016) 26:451–466
DOI 10.1007/s10516-016-9306-2

ORIGINAL PAPER

Norton and the Logic of Thought Experiments

Michael T. Stuart[1]

Received: 5 August 2016 / Accepted: 16 August 2016 / Published online: 23 August 2016

Abstract John D. Norton defends an empiricist epistemology of thought experiments, the central thesis of which is that thought experiments are nothing more than arguments. Philosophers have attempted to provide counterexamples to this claim, but they haven't convinced Norton. I will point out a more fundamental reason for reformulation that criticizes Norton's claim that a thought experiment is a good one when its underlying logical form possesses certain desirable properties. I argue that by Norton's empiricist standards, no thought experiment is ever justified in any deep sense due to the properties of its logical form. Instead, empiricists should consider again the merits of evaluating thought experiments more like laboratory experiments, and less like arguments.

Keywords Empiricism · Experimentalism · John D. Norton · Material theory of induction · Thought experiments

1 Introduction

Thought experiments seem to play a genuine role in scientific progress. However, the precise nature of that role, and how thought experiments play it, remain unclear. James R. Brown claims that some thought experiments, including Galileo's falling bodies thought experiment, can provide scientists with fallible a priori knowledge of laws of nature through direct rational perception (1991, 2). John D. Norton attempts to explain away the rationalist flavour of thought experiments by identifying them with arguments. The epistemic import of a thought experiment comes in this case from an inductive or deductive inference that only preserves or extends existing

✉ Michael T. Stuart
m.stuart@pitt.edu

[1] Center for Philosophy of Science, University of Pittsburgh, Cathedral of Learning, Room 817Q, Pittsburgh, PA 15213, USA

empirical knowledge and does not require rational insight (Norton 1991, 1996, 2004a, b). Such a defense assumes that the justificatory force of those preservations or extensions can also be reduced ultimately to experience. This latter assumption, concerning the empirical foundations of logic, has received little or no attention in the discussion of Norton's view. Yet when we focus on it, serious internal tensions appear.

In Sect. 2, I present Norton's account of thought experiments, and show that the justification of thought experiments by appeal to their logical form is in tension with Norton's own work on the nature and power of logic. In Sect. 3, I argue that when it comes to identifying the source of their epistemic value, empiricists like Norton should focus on the experimental aspect of thought experiments instead of their logical aspects.

2 The Argument View

Norton is an empiricist who claims that thought experiments are nothing more than arguments (1991, 1996, 2004a, b). It is an attractive view. After all, thought experiments are presented to us in sentences, some of which are premises, and they try to convince us of a conclusion, just like arguments. Nevertheless, most philosophers in the literature reject it. Some argue that the best explanation for the success of scientific thought experiments requires that we allow for the possibility of a direct rational grasp of relations between universals (Brown 1986, 1991, 1992, 2004), while others portray thought experiments as mental models, fictions, exemplifications, or conditions for the possibility of empirical experiments (see, e.g., Bishop 1999; Davies 2007; Gendler 2004; Gooding 1992, 1994; Miščević 2007; Nersessian 1991, 1992, 1993, 2007; Elgin 2014; Sorensen 1992; Bokulich 2001; Buzzoni 2008). Norton easily reconstructs all the counterexamples the above authors present to his view. And in any case, philosophers do admit that many thought experiments are nothing more than arguments. In other words, the above writers disagree primarily with the *scope* of Norton's claim: that all thought experiments are arguments. Despite the relative unpopularity of the view with philosophers in the literature on thought experiments, it is still very popular with philosophers in general, and continues to deserve discussion. Here are the relevant particulars of Norton's account.

According to Norton, the only "non-miraculous" way to get new information about the mind-independent world without empirical investigation is by performing logical manipulations on existing knowledge. If thought experiments produce new information about the world it is because they bring to light the hidden consequences of and relations between facts that we already know. Thus, if we wish to eliminate any appeal to "epistemic magic" (Norton 2004b, 45) we must accept that thought experiments are arguments; the only empiricist-friendly way of reasoning to new knowledge from old.

For this view to be substantial, we must be able to pinpoint what Norton means by "argument." Norton counts any inference that is licensed by a deductive, inductive, abductive or informal logic as an argument (2004b, 64), and he allows

diagrams to be steps in arguments (2004b, 58). He claims that we can be pretty sure of the presence of an argument when a piece of reasoning is convincing (1996, 353), and that the arguments underlying thought experiments will always be identifiable (1991, 129) even though they may employ tacit premises (1996, 339) and tacit inferences (1991, 142–3 note 2). Finally, the defining "mark" that tells us whether a thought experiment or argument is justified is something completely internal; it is a structural feature shared by other similarly successful inferences, something we get by "merely reading its text" (2004a, 1143). Since Norton's notion of argument is tied to his notion of logic, it is important to mention that by logic Norton means any systematization that respects the distinction between form and content. It need not preserve truth, likelihood or even be formal (2004b, 53). Arguments are things that are "governed" by such logics (2004a, 1140), and good arguments are those whose forms are approved by their logic.

This is a very broad notion of argument. There are several reasons to worry about it. First, it's not clear what it means to be "governed" by a systematization whose only feature is that it respects the distinction between form and content. Second, if being an argument depends on the state of logic, and logic evolves, then Norton is allowed to claim that the underlying argument of a thought experiment might only be revealed by future logicians (2004b, 64). This makes the impressive coincidence that thought experiments always happen to be reconstructible as arguments (Norton 1996, 339; 2004a, 1142) a good deal less impressive. This "coincidence" was supposed to motivate us to believe that all thought experiments are arguments, but if Norton's claim includes the additional premise that all good reasoning will eventually be formulated by future logicians, this means that there can be no true counterexamples to his hypothesis until we have reached the end of logic. Or worse, it defines all possible counterexamples away as instances of bad reasoning. Another way to put this is that every inference we count as successful will necessarily be an argument since we are assuming that logic will eventually get around to formalizing all successful inference forms. I believe this is behind rejoinders like the following from Brown: "Norton says that thought experiments are often *disguised*, not explicit arguments. So the real claim is that they can be reconstructed along his empiricist lines. Existence claims like this are devilishly difficult to defeat. I doubt that an actual refutation could ever be delivered" (Brown 1992, 275).

These charges deserve consideration. For the purposes of this paper, however, I will assume some non-trivial notion of argument can be specified so we can move on to the more fundamental issue, which concerns the claim that we can justify cognitive activities that produce epistemologically desirable changes in our overall doxastic states by reference to the logical properties of those activities (or propositional reconstructions of those activities). This obviates the need to discuss arguments by focusing on the source of their justification: approved logical form. Thus, whether the set of cognitive activities we call arguments and those we call thought experiments are actually identical doesn't matter for present purposes, although it does for Norton's account. I want to focus on the more fundamental assumption that drives Norton's view: that we can justify the output of thought experiments by reference to the logical properties of the inferences that make up those thought experiments.

A natural way to begin an inquiry into empiricist accounts of logical justification is by distinguishing deductive and inductive justification. I begin with deduction, about which Norton writes: "Deductive inferences merely restate what we have already presumed or learned. There is no mystery in what powers the inference and permits the conclusion. We are just restating what we already have in the premises. The warrant lies fully within the premises. If we know all winters are snowy, it follows deductively that some winters are snowy" (Norton forthcoming, Chapter 2, 7–8).

2.1 Deductive Thought Experiments

Let's take a case. Galileo invents a now-infamous thought experiment to criticize the Aristotelian theory of motion according to which heavier bodies fall faster than light ones. Here is Norton's reconstruction (1996, 341–342):

1. Assumption for *reductio* proof: The speed of fall of bodies in a given medium is proportionate to their weights.
2. From 1: If a large stone falls with 8 degrees of speed, a smaller stone half its weight will fall with 4 degrees of speed.
3. Assumption: If a slower falling stone is connected to a faster falling stone, the slower will retard the faster and the faster speed the slower.
4. From 3: If the two stones of 2 are connected, their composite will fall slower than 8 degrees of speed.
5. Assumption: the composite of the two weights has greater weight than the larger.
6. From 1 and 5: The composite will fall faster than 8 degrees.
7. Conclusions 4 and 6 contradict.
8. Therefore, we must reject Assumption 1.
9. Therefore, all stones fall alike.

Gendler (1998) provides excellent criticism of this reconstruction. But for the sake of argument let's assume that it is both historically faithful to Galileo and his contemporaries, as well as logically valid. Why should we believe its conclusion? Since the mark that tells us whether a thought experiment is justified "is *just* that the thought experiment either uses an argument form licensed by a logic or can be reconstructed as one" (2004b, 54; my emphasis), and since we're assuming this thought experiment can be reconstructed as a reductio ad absurdum, we should conclude that it is a good thought experiment.

But validity is not soundness; we can still deny a premise in the reconstruction of the thought experiment. This is an obvious point, but it means that Norton's claim that a thought experiment is justified "just" when it can be reconstructed as a good (valid, cogent, etc.) argument is actually the claim that a thought experiment is justified when it has true (or likely) premises *and* can be reconstructed as a good argument. This means that reference to the logical features of a thought experiment's reconstruction is never going to provide *all* of a thought experiment's justification, even for Norton. And so in the remainder of this paper I will focus on

the weaker claim that reference to the logical features of a thought experiment can provide *some* of its justification. The new question is: do deductive thought experiments provide some (any) justification for their conclusions that is not already to be found in the premises and our background beliefs?

It is important that we look closely at what Norton says, since the quotation with which I ended the last section claims that the warrant for a deductive conclusion "lies fully within the premises," yet Norton claims that "the mark [of a good thought experiment] is just that the thought experiment either uses an argument form licensed by a logic or can be reconstructed as one" (2004b, 54). This latter claim says that possessing a certain logical mark *just is* what justifies a thought experiment. How do we navigate between these two claims? The natural thing to do is deny this apparent dichotomy: what makes a deductive conclusion justified is having true premises and instantiating an approved logical schema. Such an interpretation goes against some of Norton's more radical remarks. For example, Norton writes that inferring with a logical license is justified because there is "something in the logic that evidently confers the power of a thought experiment to justify its conclusion" (Norton 2004b, 54). Interpreting statements like these so that logic confers only *part* of the total justification (to respect the point above concerning soundness), we now want to know what it is about instantiating an approved logical form that adds justification to an inference.

Reconstructing a thought experiment as an argument allows us to see whether it is of an approved sort. But this act of checking a reconstruction against a list of approved logical schemas is only justification by categorization, or justification by proxy. Reconstructing a thought experiment as an instance of modus ponens is nothing more than saying that it is relevantly similar to other inferences that we also take to be justified. And this is not what we want from a normative theory. And Norton seems to agree: "the mark [of a good thought experiment] must be something that we can recognize *in the thought experiment itself*. It cannot merely be that the thought experiment is found in an approved list; that would be an external mark" (2004a, 1143; my emphasis). If resulting from the use of a deductively valid schema is a fact that adds justification, we must find something about that schema that justifies using it. The fact that such schemas help to categorize inferences that we already recognize as justified is not enough. With this we decide the logical Euthyphro problem: Logical forms are not good because logicians say they are good. Logicians say they are good because they are good.

But what about a good logical form tells a logician that it is good? For an empiricist, it cannot be that the good-making properties of abstract logical forms exist and are "seen" in the rationalist's sense. If instantiating an approved logical form is going to add justificatory power, therefore, we need logical approval to be more than what logicians happen to like, and less than direct perception of features of abstract logical forms. Two empiricist strategies (taken from philosophy of mathematics) might be useful here. An empiricist can portray logical form as real only in the sense of 1) relations abstracted from experience or 2) as useful tautologies.

Let's begin with the first strategy, according to which logical form is abstracted from experience (and logical approval attaches to features of those abstractions).

Approval of logical form in this sense could not provide the ultimate justification for inferences, precisely because such form arises as a representation of something non-formal which must already be approved. For example, we discover that inferences of a certain type are successful. Logicians abstract their logical form and approve of it. But approval of the abstract reconstruction of good inferences is only justificatory in so far as it is shorthand for approval of the inferences themselves (I take this to be a form of the point made in Carroll 1895). Approval of logical form (where logical form is thought of as a set of relations abstracted from experience) is again only approval by proxy, and can therefore be justification only by proxy. Properties of logical form thus conceived cannot serve as the "internal mark" of a good thought experiment (or argument) at a fundamental level.

The other empiricist strategy listed above portrays logical form as a system of conventional but useful tautological relations, "a convenient fiction that is very helpful for purposes of calculation, and helpful too as providing a vocabulary with which to express our scientific theories" (Bostock 2009, 226). This quotation concerns empiricist accounts of mathematics, but the point translates. Tautological systems like mathematics and logic are useful because they allow us to make extremely subtle distinctions in our concepts, draw new inferences, display the consequences of our commitments, point out contradictions, and so on. However, when we conceive of mathematics and logic as conventional languages, we face the problem that such languages only ever express and never justify truths about the non-linguistic world.

In this sense, approved logical form would again be unable to provide any of the ultimate justification for the conclusion of a thought experiment. Of course, the same conclusion holds *mutatis mutandis* for empiricist accounts of mathematics. An empiricist who claimed that our knowledge of a real-world system was justified in a fundamental sense by that system's instantiating a certain kind of approved mathematical form would be estranging her own empiricism.

Returning to the question we asked above: despite his remarks to the contrary, perhaps the best interpretation of Norton's view would claim that instantiating approved deductive logical form *does not* confer any additional justificatory power. That is, no new justification is provided to a proposition by being the result of a sound deduction. Norton's account would therefore imply that we never gain new knowledge from a thought experiment that is deductive. This is somewhat surprising, however, given Norton's deductive reconstructions of many of the most powerful thought experiments in science, including the one above by Galileo, Einstein's magnet and conductor (1991, 135–136), Einstein's elevator (1991, 136–138), Einstein's two fluid bodies (1991, 138–139), Einstein's falling bodies thought experiment (1993, 6–8), Einstein's spinning top (1993, 8–12), Einstein's stressed rod (1993, 12–15), Einstein's radiation in a mirrored box (1991, 133–134; 1993, 15–19), Einstein's lowering and raising a stressed rod (1993, 19–24), Einstein's moving rod and slot thought experiment in special relativity (2004a, 1145), Einstein's rotating disk (2004b, 50–51), and Einstein's clock in the box (2004b, 63–64).

Either none of these thought experiments increase our knowledge, or despite their deductive reconstructions by Norton, they are inductive. Norton admits that some of

them do provide knowledge, since he grants it in the case of Einstein's mirrored box (1993, 17), Einstein's rod and the slot (2014, 1148) and he allows in several places that thought experiments can produce knowledge (2004a, 1140; 2004b, 44, 49–50) justification (1991, 142; 1996, 339, 354; 2004a, 1143; 2004b, 53–54) and new information about the empirical world (1991, 129; 1996, 333). So perhaps we should conclude that despite their deductive-looking reconstructions, the logical form that enables these thought experiments to produce new knowledge is actually inductive. And we should then say in general that any thought experiment that increases our knowledge is inductive.

Once we turn to inductive thought experiments, however, the plot thickens.

2.2 Inductive Thought Experiments

Norton is explicit that his account is intended to capture inductive as well as deductive thought experiments (Norton 2004b, 64; see also 49). However, it is not clear that inductive inferences *have* logical form, and this is something emphasized by Norton himself. In recent years Norton has developed a new theory of induction (covering abduction, informal and probabilistic reasoning) according to which inductive arguments *do not admit of a distinction between form and content*. Norton's new theory of inductive inference is called the "material theory of induction" (Norton 2003, 2005, 2010, 2011, 2014, forthcoming), which is meant as an alternative to formal theories of induction. Formal theories of induction attempt to provide rules of inductive inference, whether in natural language or using a formal calculus such as Bayesian probability theory. These attempts separate the material of an inductive inference from its form, in order to isolate what is common to the form of successful inductive inferences. According to formal theories of induction, "valid inductive inferences are distinguished by their conformity to universal templates. They may be simple, such as the template that licenses an inference from some past A's being B to the conclusion that all A's are B. Or they may be more complicated, such as the requirement that degrees of inductive support conform to the probability calculus" (Norton 2014, 673).

The endeavour to find the approved logical schemas for good inferences is precisely what Norton applauded earlier, and now rejects for induction: "If one adopts a material theory of induction, one no longer separates factual content from the rules of inductive inference. The problem of justifying some particular induction is replaced by the straightforward task of justifying the facts that warrant it" (2014, 672).

If we stop separating the rules of inference from the "factual content" as Norton recommends, then we eliminate the distinction between the form and the content of inductive inferences. This means that according to Norton's own definition of a good argument, which requires "the familiar distinction between the form and content" and instantiating "an argument form licensed by a logic" (2004b, 53), inductive arguments are not arguments, and inductive thought experiments are not either. *A fortiori*, inductive thought experiments will not be individuated or evaluated based on their formal features, since they do not admit of a form/content distinction that can be used by logic to identify good inferences.

Norton argues that individual inductive inferences do follow *local* rules, so he might try to escape this criticism by saying that the mark of a good inductive thought experiment is to be found in the local logical form of the reconstructed inductive argument. This won't do, however, since the power of those local forms must still ultimately derive from facts concerning the material invoked by the inductive inference. It would be strange if Norton's material theory of induction was only material at the most general level, and fell back on formal inductive theories everywhere else, since the heart of the material theory of induction is the claim that, "what separates the good from the bad inductive inferences are background facts, the *matter* of the inference, as opposed to its *form*" (forthcoming, chapter one, page 4). The rule "fire causes smoke" locally justifies the inference from "there is smoke" to "there is fire." That is, "fire causes smoke" is a rule that is only applicable to a certain (local) class of phenomena. Still, this rule is justified by and inseparable from facts about smoke and fire, for Norton. We can appeal to the rule that fire causes smoke as a justification for our alarm when seeing a great deal of smoke, given that we find ourselves in a domain in which that rule applies (there are flammable objects as well as oxygen and the possibility of sparks or heat). But when we ask what justifies *that* rule, we cannot answer with reference to any of the purely formal properties of our inference. Perhaps we could appeal to some formal properties of fire or smoke themselves, but those formal properties are not what justify our inference from smoke to fire. It's the causal properties of fire and smoke we learn through experience (and testimony) that do this, not formal ones. To repeat, our experience with the material-causal relations is what justifies our ability to make inductive inferences using them, not any logical-formal properties those relations may have.

There is therefore a serious internal tension between Norton's account of thought experiments according to which thought experiments are filled with irrelevant but picturesque details on the one hand, and his account of induction according to which the particular details are crucially important for justification, on the other. Having considered both deduction and induction, then, an empiricist like Norton cannot consistently hold the following three claims: that thought experiments provide knowledge, that thought experiments are justified by properties of their logical forms, and that induction is material.

This is because if thought experiments provide knowledge, they are inductive. And if inductions are justified by the properties of their logical forms, the material theory of induction is mistaken. Next, if thought experiments provide knowledge and the material theory of induction is correct, this contradicts the heart of Norton's argument view: that thought experiments receive their ultimate justification from appeal to their logical-formal properties. Finally, if thought experiments are justified by properties of their logical forms and the material theory of induction is correct, then no thought experiment provides knowledge, and Norton claims in many places that they do.

In sum, Norton must give up the claim that thought experiments provide knowledge, or his account of thought experimental justification, or his material theory of induction.

With that I conclude my criticism of Norton's view that thought experiments are arguments. Since my criticisms have been specific to the combination of views held by Norton, it would be instructive to consider whether another empiricist or naturalist account of thought experiments can be created that explains their ultimate source of justification in a way that is consistent with the empiricism of someone like Norton.

If someone was tempted to go in this direction, they might start by going back to the basic tenants of empiricism to decide which of the three claims to reject. The fundamental source of justification for empiricists like Norton are facts about experience. So perhaps it is facts about experience that justify inferences. And indeed this is precisely what Norton is committed to with respect to his material theory of induction. All an empiricist needs to do is extend the material theory of inference to thought experiments, which, if we accept that they can provide new knowledge when they are inductive, requires us to leave behind the commitment to logical form as a justifying mark on the fundamental level.

To see how this might go, let's consider a simple example. We ask someone whether square objects can fit through circular holes. This person cuts a square of paper and drops it into our mug of coffee, annoyed by the absurdity of our question. She has answered by direct reasoning. Alternatively, she could *imagine* cutting a square of paper and dropping it into our coffee mug, *infer* that there is nothing to prevent her from doing this, and provide the following report: "I could fit a square object through a circular hole, and if I can do that, then there are at least some square objects that fit through circular holes. So yes, square objects can fit through circular holes." Consider one final strategy, where the subject performs the same mental actions as in the second case, but then also presents her reasoning in a formally reconstructed manner.

In the first case, what made our subject's action possible was the absence of any physical law-like connection between the areas of everyday squares and the diameters of everyday circles. *This exact same fact* is what justifies the argument provided in the words of the second case, and the symbols of the third case. At least, this is what an empiricist should claim.

The same is true of inductive thought experiments and arguments. Suppose we could and did reconstruct all inductive thought experiments as arguments. Because the same facts about experience are what justify thought experiments and arguments, they will always enjoy exactly the same degree of justification. This would satisfy both Norton's "reconstruction" and "reliability" theses: that all thought experiments are reconstructible as arguments and that all thought experiments are equally justified as their reconstructed arguments (e.g., 1996, 339 and 2004b, 52 respectively). These are supposed to be surprising facts that warrant the equation of thought experiments and arguments. However as we have seen, even if both of these were true, we would still have no reason to believe that a given thought experiment was justified by appeal to its reconstructed argument, or that it was identical to its reconstructed argument.

In sum, if facts about experience are the ultimate ground for knowledge, and if thought experiments sometimes produce new knowledge, then we should look at the various ways we can generate new facts from old experience through thought.

Perhaps the most epistemologically interesting and relevant of these is the running and interpretation of experiments. So let us turn to thought experiments both as interpretations of experiments, and as experiments themselves.[1]

3 Material (As Opposed to Formal) Accounts of Thought Experiments

Experiments begin with a question concerning a system of interest, proceed through a controlled intervention on or observation of that system (or a representation of that system), and end with an interpretation of what transpired that hopefully addresses our original question. To emerge with a high degree of confidence in the conclusion of an experiment, we sometimes discuss our confidence in each of these stages: we want to be confident that our question is the right one, that our representations are accurate, that our manipulations target the right parts of the system, that our observations are precise, and that our interpretations are free of bias.

If this is correct, then perhaps the ultimate justification of a thought experiment that produces new knowledge lies with whatever makes experiments and their interpretations reliable in general. Let's start with five criteria for a good experiment, drawn from the classic Franklin (1986).

1. The system to be investigated must be well-isolated, abstracting away irrelevant features of the surrounding environment.
2. Experimental bias must be eliminated.
3. Sources of error must be understood and accounted for.
4. Instruments should be calibrated as well as possible.
5. We should have a theory of our instruments.

These criteria may be applied to thought experiments where we understand the instrument of observation to be the mind, broadly construed. Here is an extremely brief sketch of how this might go.

1. We can isolate the relevant features of interesting systems in our imaginations through abstraction and idealization. Isolation of properties and processes in the imagination is easier to do than the material equivalent, and this is likely one reason thought experiments are so ubiquitous: they are economical when it comes to this first requirement. There is a large and rapidly expanding literature on how we perform these tasks (e.g., Giere 1988; Godfrey-Smith 2009; Laymon 1980; Matthews 2004; Portides 2011; Weisberg 2007).
2. We must eliminate experimental bias. In the case of thought experiments, experimental bias is cognitive bias, and especially confirmation bias, which is a fairly well-understood phenomenon. One way to keep confirmation bias in check is by discussing thought experiments with others who do not share our

[1] Several other philosophers have discussed the relationship between thought experiments and experiments more generally, especially Buzzoni (2008), Gooding (1992, 1994) and Sorensen (1992). But none have tried to make this connection within the context of an empiricist account of thought experiments. While I am not committed to empiricism, I think this is an important and interesting project, and an extension of Nersessian's "empiricism without logic" (Nersessian 2007).

expectations, and this is partially what we see in classrooms, seminars, conferences and journals, and confirmed as a desideratum by the growing corpus in experimental philosophy. We have learned that as much as possible we should present imaginary scenarios in different words, orders and using different contents, to isolate for priming effects.

3. There are many sources of error in thought experiments, including inaccurate representations of the target system and weak imaginations. These problems must be appreciated, but they are not reasons to be skeptical about thought experiments in general. Laboratory experiments can also fail when their representations are inaccurate, or when their instruments have powers of resolution that are too weak to yield a single conclusion on interpretation. One might object that laboratory experiments have better ways of checking their representations for accuracy, but if this is true it seems to be a matter of degree, as coherence is our main check in both cases: laboratory experiments require coherence between theories, experience and across experimental repetition. Thought experiments require coherence with all of our knowledge, including that which derives from experience, as well as coherence across repetition by dissenting scientists (Popper 1935, 464–480). This is not to say that coherence guarantees success, merely that the method for error checking is relevantly similar in both cases.
4. Our imaginations must be calibrated. And they are, through years of practice with hypothetical reasoning. In all skilled positions from electrician to tennis player, we will have considered abstract cases almost every day as part of our education and work. This is not to say that thought experiments are always justified when they are performed by someone who has spent years working in the subject area of the thought experiment. For example, the recent objections against the use of *philosophical* imagination might be more worrying than the same objection applied to the use of thought experiments in other fields like engineering or mathematics (Thagard 2014, and see Buzzoni, forthcoming). Still, we care about philosophical intuitions. Here are two reasons for optimism. First, there are methods for resolving conflict. This is again the same as in the case of laboratory experiments. If two well-calibrated experiments provide robust and conflicting outcomes, the reaction is either to make a theoretical change or perform more experiments. If the theoretical change is not ad hoc, it will have empirical consequences that can be tested. It is therefore ultimately to experience that we make recourse. Second, intuitions in philosophy and science work in similar ways—mostly as hypotheses. In this case, they do not bear evidential weight, and in fact they can be quite mistaken and still lead eventually to epistemologically warranted outcomes (Buzzoni, forthcoming, also argues that scientific and philosophical thought experiments will be evaluated in the same way, despite their different "directions of fit" to the world. For more discussion on philosophical reasoning as experimental see Stuart 2015).
5. To have a theory of our instrument we need a theory of inference-making. Much work in this direction has been carried out in psychology, sociology, phenomenology, philosophy of mind, neuroscience, and more broadly speaking,

the arts and humanities. It would be difficult to deny that these projects have provided any insights into how we make (good) inferences. Studies in logic, informal reasoning, and work on cognitive biases alone represent important strides in the last century.

In sum, whether thought experiments are reconstructed as arguments or not, applying the norms of laboratory practice seems prima facie to be a good way to evaluate their level of justification for an empiricist. And there are myriad resources from which to build such an account, not least including cognitive science, which recognizes and explains how to avoid some of the pitfalls of using imaginary scenarios in certain contexts.

I leave the full explication of such an account for future work. For now, I want to deal with two objections concerning the *differences* between thought experiments and laboratory experiments. The first claims that thought experiments cannot intervene on extra-mental systems, which is one of the most important features of a laboratory experiment. Here are two replies to such an objection.

Intervention—while important—is not a necessary condition for experimentation. There are many examples in science where an experiment does not intervene. A classic example concerns peppered moth camouflage (Kettlewell 1955, 1956, 1958). There were several colourings of peppered moths in England, and it appeared that the darker (and rarer) forms were becoming more common in areas downwind of factories responsible for industrial pollution that darkened the trees and killed much of the lichen cover. To see if the spread of the melanic gene was due to the inability of predators to spot the darker moths against the darkened trees, Henry Kettlewell collected moths of different colours and placed them on different coloured tree barks in a laboratory setting. Predators were then brought into see which moths were preyed upon. Kettlewell found a significant advantage for the darker moths when darkened tree bark was used: they were more than twice as likely as the paler moths to avoid predation. This was well-prepared *observation* of natural processes, not intervention. Nevertheless, it has all the trappings of a controlled experiment, and it established a historically important result.

Another example is Pasteur's famous set of experiments in which he showed that spontaneous generation (of insects, rats, etc.) does not occur when soiled rags or dead flesh are kept away from circulating air. He established this result simply by watching rags and meat kept under limited air circulation (see Conant 1953). Again, we shouldn't describe Pasteur's experiments as interventions, because we wouldn't be able to say *on what* he intervened. The life cycles of hypothetical houseflies that might have lived if their progenitors had been able to lay eggs on a particular piece of meat hardly forms a system on which we could intervene. Like Kettlewell, what Pasteur did was create a new and different system, one in which there were no housefly eggs, and observed it, finding that a state without houseflies begets no houseflies.

The second response to the charge that thought experiments cannot intervene is to affirm that thought experiments *can* intervene, by making an analogy to other homomorphic representations (like material models). When a scientist aims to learn about a given natural system, she primarily does her intervening on idealized

versions of these systems. We can map the strengths of different representational relations on a continuum that runs from direct material reasoning (like putting square objects through circular openings), to scientific field experiment, laboratory experiment, material modelling, computer simulation, thought experiment, all the way to reasoning with representations that have almost nothing in common with what they represent (like the dots and dashes of Morse code, or binary representations in a computer). When we consider this continuum, we notice that *direct intervention* is an ideal that typically goes unachieved. Almost always, we intervene on abstracted and idealized versions of natural systems, not on the systems themselves. And thought experimental scenarios in science are abstracted or idealized versions of natural systems. It is true that we cannot be sure that a mentally represented system will react as a natural system would, but this problem is only gradually more severe than it is for material models, computer simulations and laboratory experiments. In all these cases, we hope that our representations of natural systems preserve enough of the relevant structure of the natural systems that something new can be learned from them. This hope is grounded in the quality of our assumptions, confidence in our background theory, and understanding of our instruments.

But how do we *intervene* on a representation? To intervene experimentally is to interrupt or otherwise tamper with the time-evolution of a process. When done properly, we are able to gauge counterfactually how the trajectory of the process compares pre- and post-intervention using a control. There are several authors who point out the dynamic aspects of thought experiments (Miščević 1992, 220; Nersessian 2007, 143, and forthcoming). According to an interpretation of thought experiments which portrays them as dynamic, we can take the control case to be the evolution of the system according to the rules and constraints we build into the scenario without an intervention. The rules and constraints that determine the time-evolution of a system come partially from representational conventions, partially from background knowledge. Just as with all systems of representation, there are conventions governing acceptable sets of outcomes for imagined scenarios. The constraints in a scientific thought experiment will often be more flexible than the constraints on say, Euler circles or Venn diagrams, but this is not to say they don't exist. Looking at the way we reason with fiction might reveal some general types of constraint (Ichikawa and Jarvis 2009), and common sense tells us that particular scenario types will have specific rules that vary with the content (for example, Sherlock Holmes can only do what a human can do). Given that there are rules and constraints that we use to imagine the evolution of a system, there will be interventions we can perform on the control case by altering the initial conditions or the constraints themselves, and then "seeing" how it plays out (see Nersessian, forthcoming). And this is exactly what we find in cognitive science studies of how students create and interact with thought experiments in the classroom (Kösem and Özdemir 2014).

Finally, we might worry that the imagined results of a thought experiment thus conceived will be hopelessly indeterminate compared to a laboratory experiment since we can imagine any outcomes to a given scenario that we want, and real-world systems only evolve in only one way. But this is not necessarily true. Real-world

systems often have as many or more degrees of freedom than a thought experiment does—consider all the possible outcomes of a social or psychological experiment, for instance. And we have greater control over the constraints involved in a thought experiment, which we can tighten as we see fit. In any case, the real concern is whether the constraints we place in our imaginations for a given scenario reliably track the constraints present in a system of interest, to some degree. The answer to this question will depend on the system, the quality of our imagined scenario, and our training; it cannot be determined in advance.

4 Conclusion

Norton argues that thought experiments are arguments. According to Norton, arguments are inferences governed by logic, and their conclusions are justified by logic. For an empiricist, logical justification must boil down to empirical justification. According to Norton, deduction merely rearranges what we already know, but it does not provide any new justification. And according to Norton's account of induction, inductive inferences are justified because of the real-world relations between the material of the premises of the argument, and not their logical properties.

Therefore, deductive thought experiments say nothing new, and inductive thought experiments are not justified by appeal to logic. If new knowledge results from a thought experiment, which Norton allows, then Norton's account is incapable of explaining it by reference to features of logical form. This is problematic because it is a desideratum of any epistemological account of thought experiments that it provide some explanation of the source of thought experimental justification.

That said, we are not about to stop representing thought experiments logically as arguments; doing so is a helpful clarificatory exercise that provides a quick means of separating good from bad inferences. But for a complete epistemology of thought experiments, we need more than handy heuristics. We must identify the fount of justification, and for an empiricist, reference to logical form will not do. This seemingly minor point is crucial when applied to the literature on thought experiments, because with it, we remove a fundamental component of Norton's account, which has been very influential in the literature.

Empiricists may still claim that thought experiments are regular inferences that are justified when that particular sort of inference is justified, but they will need a story about how those inferences themselves are justified, and it will not proceed by appeal to logical schemas if we accept the material theory of induction. In light of this problem, I have suggested an alternative source for a thought experiment's justification: experimental interaction and interpretation, whose epistemic evaluation would follow similar criteria as in other sorts of experiments and interpretations. This might be a preferable direction for empiricists in the ongoing discussion of the epistemology of thought experiments, although it is equally open to naturalists and rationalists as well.

Acknowledgments I'd like to thank John D. Norton, Nancy Nersessian, Yiftach Fehige, James R. Brown, Marco Buzzoni, Joseph Berkovitz, Catherine Elgin, Sören Häggqvist, Elke Brendel, Geordie McComb and an anonymous referee for comments and discussion, as well as audiences at the University of Toronto, Bonn and Pittsburgh. This research was supported by an Ontario Graduate Research scholarship, the Germany/Europe fund from the University of Toronto, and a postdoctoral fellowship at the Center for Philosophy of Science at the University of Pittsburgh.

References

Bishop M (1999) Why thought experiments are not arguments. Philos Sci 66:534–541
Bokulich A (2001) Rethinking thought experiments. Perspect Sci 9:285–307
Bostock D (2009) Empiricism in the philosophy of mathematics. In: Irvine A (ed) Handbook of the philosophy of science: philosophy of mathematics. Elsevier B.V, Amsterdam, pp 157–229
Brown J (1986) Thought experiments since the scientific revolution. Int Stud Philos Sci 1:1–15
Brown J (1991)[2011] The laboratory of the mind: thought experiments in the natural sciences. Routledge, London
Brown J (1992) Why empiricism won't work. Proc Philos Sci Assoc 2:271–279
Brown J (2004) Why thought experiments do transcend empiricism. In: Hitchcock C (ed) Contemporary debates in the philosophy of science. Blackwell, Malden, pp 23–43
Buzzoni M (2008) Thought experiment in the natural sciences. Königshausen & Neumann, Würzburg
Buzzoni M (forthcoming) Thought experiments in philosophy. Topoi
Carroll L (1895) What the tortoise said to achilles. Mind 4:278–280
Conant J (ed) (1953) Pasteur's and Tyndall's study of spontaneous generation. Harvard University Press, Cambridge
Davies D (2007) Thought experiments and fictional narratives. Croat J Philos 7:29–45
Elgin C (2014) Fiction as thought experiment. Perspect Sci 22:221–241
Franklin A (1986) The neglect of experiment. Cambridge University Press, Cambridge
Gendler T-S (1998) Galileo and the indispensability of scientific thought experiment. Br J Philos Sci 49:397–424
Gendler T-S (2004) Thought experiments rethought—and reperceived. Philos Sci 71:1152–1163
Giere RN (1988) Explaining science: a cognitive approach. The University of Chicago Press, Chicago
Godfrey-Smith P (2009) Models and fictions in science. Philos Stud 143:101–116
Gooding D (1992) What is experimental about thought experiments? PSA: Proc Bienn Meet Philos Sci Assoc 2:280–290
Gooding D (1994) Imaginary science. Br J Philos Sci 45:1029–1045
Ichikawa J, Jarvis B (2009) Thought-experiment intuitions and truth in fiction. Philos Stud 142:221–246
Kettlewell HB (1955) Selection experiments on industrial melanism in the *Lepidoptera*. Heredity 9:323–342
Kettlewell HB (1956) Further selection experiments on industrial melanism in the *Lepidoptera*. Heredity 10:287–301
Kettlewell HB (1958) A survey of the frequencies of *Biston betularia* (L.) (Lep.) and its melanic forms in Great Britain. Heredity 12:51–72
Kösem Ş, Özdemir Ö (2014) The nature and function of thought experiments in solving conceptual problems. Sci Educ 23:865–895
Laymon R (1980) Idealisation, explanation, and confirmation. In: Asquith PD, Giere RN (eds) PSA 1982, vol 1. Philosophy of Science Association, East Lansing, pp 336–350
Matthews RM (2004) Idealisation and Galileo's pendulum discoveries: historical, philosophical and pedagogical considerations. Sci Educ 13:689–715
Miščević N (1992) Mental models and thought experiments. Int Stud Philos Sci 6:215–226
Miščević N (2007) Modelling intuitions and thought experiments. Croat J Philos 7:181–214
Nersessian N (1991) Why do thought experiments work? Proc Cogn Sci Soc 13:430–438
Nersessian N (1992) How do scientists think? Capturing the dynamics of conceptual change in science. In: Giere RN (ed) Cognitive models of science. University of Minnesota Press, Minneapolis, pp 3–44
Nersessian N (1993) In the Theoretician's laboratory: thought experimenting as mental modeling. Proc Philos Sci Assoc 2:291–301

Nersessian N (2007) Thought experiments as mental modelling: empiricism without logic. Croat J Philos 7:125–161
Nersessian N (forthcoming) Cognitive science, mental modelling, and thought experiments. In: Stuart et al. (eds) The routledge companion to thought experiments. Routledge, London
Norton J (1991) Thought experiments in Einstein's work. In: Horowitz T, Massey G (eds) Thought experiments in science and philosophy. Rowman & Littlefield, Lanham, pp 129–148
Norton J (1993) Einstein and Nordstrom: some lesser-known thought experiments in gravitation. In: Earman J, Janssen M, Norton JD (eds) The attraction of gravitation: new studies in the history of general relativity. Birkhauser, Boston, pp 3–28
Norton J (1996) Are thought experiments just what you thought? Can J Philos 26:333–366
Norton J (2003) A material theory of induction. Philos Sci 70:647–670
Norton J (2004a) On thought experiments: is there more to the argument? Philos Sci 71:1139–1151
Norton J (2004b) Why thought experiments do not transcend empiricism. In: Hitchcock C (ed) Contemporary debates in the philosophy of science. Wiley-Blackwell, Somerset, pp 44–66
Norton J (2005) A little survey of induction. In: Achinstein P (ed) Scientific evidence: philosophical theories and applications. Johns Hopkins University Press, Baltimore, pp 9–34
Norton J (2010) There are no universal rules for induction. Philos Sci 77:765–777
Norton J (2011) History of science and the material theory of induction: Einstein's Quanta, Mercury's Perihelion. Eur J Philos Sci 1:3–27
Norton J (2014) A material dissolution of the problem of induction. Synthese 191:671–690
Norton J (Forthcoming) The material theory of induction. http://www.pitt.edu/∼jdnorton/homepage/cv.html#material_theory
Popper K (1935)[2005] The logic of scientific discovery. Routledge, London
Portides D (2011) Seeking representations of phenomena: phenomenological models. Stud Hist Philos Sci Part A 42:334–341
Sorensen R (1992) Thought experiments. Oxford University Press, Oxford
Stuart M (2015) Philosophical conceptual analysis as an experimental method. In: Gamerschlag T, Gerland D, Osswald R, Petersen W (eds) Meaning, frames and conceptual representation. Düsseldorf University Press, Düsseldorf, pp 267–292
Thagard P (2014) Thought experiments considered harmful. Perspect Sci 22:288–305
Weisberg M (2007) Three kinds of idealization. J Philos 104(12):639–659