Thought Experiments as Tools of Theory Clarification

GRACE HELTON

Draft of June 2023
Where possible, please cite final version in Routledge Studies in Epistemology

Key words. intuition, thought experiment, method of cases, metaphilosophy, understanding, disconfirmation, theory clarification, epistemology, metaethics, philosophy of science

Word Count. 7,536 (inclusive of notes); 6,606 (exclusive of notes)

Abstract. It is widely presumed that intuitions about thought experiments can help overturn philosophical theories. It is also widely presumed, albeit implicitly, that if thought experiments play any epistemic role in overturning philosophical theories, it is via intuition. In this paper, I argue for a different, neglected epistemic role of philosophical thought experiments, that of improving some reasoner's appreciation both of what a theory's predictions consist in and of how those predictions tie to elements of the theory. I call this role *theory clarification*. I show that theory clarification does not proceed via intuition, and I argue that it is only in conjunction with theory clarification that intuitions about thought experiments can help overturn philosophical theories. I close by sketching how a more radical view might be true, on which thought experiments help justify the rejection of philosophical theories *exclusively* by clarifying theories, not by any intuitions those thought experiments might generate.

Among contemporary philosophers, it is a pervasive view that *intuitions* about thought experiments can sometimes help justify the rejection of relevant philosophical theories. It is controversial whether intuitions are judgments, dispositions to judge, *sui generis* mental states, or some other kind of attitude.[1] For present purposes, we can characterize intuitions as, minimally, mental states which are generated in a non-inferential fashion and which have a certain phenomenal quality, or 'feel.'[2] In the relevant sense, *thought experiments* (sometimes called *cases*) are scenarios or mental models of scenarios, typically introduced via a vignette.[3]

An example will help illustrate the mainstream view about how intuitions about cases can help overturn philosophical theories. Consider *act utilitarianism*, the ethical view according

[1] For the belief view, see Gopnik & Schwitzgebel (1998) and Ludwig (2007); For the disposition to believe view, see Pust (2000) and Erlenbaugh & Molyneux (2009). For the *sui generis* state view, see, Bengson (2015), Brogaard (2014), Chudnoff (2011), and Koksvik (2011).
[2] Even the intuition skeptic Cappelen (2012) and the intuition advocate Brogaard (2014) can agree on this much. For a helpful overview, see Pust (2019).
[3] See Brown & Yiftach (2022) for a helpful overview of the nature of thought experiments.

to which some action is morally permitted just in case it maximizes well-being. Utilitarianism is an initially appealing view and as anyone who has taught an introductory ethics course can attest, it is a view whose appeal to the novice philosopher can seem undeniable.

One way of criticizing utilitarianism appeals to thought experiments, such as the following:

> TRANSPLANT CASE
> A doctor oversees six patients, five of whom will die if they do not quickly receive organ donations. The sixth patient has organs which could be distributed to the other five to save all five of them. Without the consent of this sixth patient, the doctor kills him and distributes his organs to the others, thus saving five people.[4]

Many people respond to the doctor's action with disapproval, even horror. They think that in killing the sixth patient, she does something morally wrong, even though her action saves several lives. Drawing on this intuition, critics of act utilitarianism develop the following argument:

> **Argument Against Act Utilitarianism**
> (1) The doctor's action in the transplant case is morally wrong.[5]
> (2) The doctor's action (the same action as in (1)) maximizes well-being.
> (3) According to act utilitarianism, the doctor's action in the transplant case is not morally wrong.
> (4) Act utilitarianism is false.

Most ethicists—utilitarians and non-utilitarians alike—take this style of argument to provide a powerful challenge to utilitarianism. But how exactly does the thought experiment, understood as the scenario about the doctor, relate to the argument? The mainstream view goes something like this: The thought experiment elicits a powerful intuition that the doctor's action is morally wrong. This intuition is itself evidence for or else helps generate evidence for (1), the claim that the doctor's action is morally wrong. (1) in turn combines with other claims about the doctor's action, namely that it is one which maximizes well-being and one which is therefore sanctioned by utilitarianism, to help buttress the overall conclusion that utilitarianism is false.[6]

---

[4] Thomson (1976).
[5] It is contentious whether (1) should be interpreted as a necessity claim, a counterfactual claim, a possibility claim, or something else. See Williamson (2007), Ichikawa & Jarvis (2009), Malmgren (2011), and Pust (2019) for discussion and competing perspectives.
[6] Often, the method of cases is construed as part of a broader a strategy of moving between less general and more general considerations. In this paper, I'm neutral on how less general considerations, such as those

On this picture, the intuition about the wrongness of the doctor's action plays an epistemic role in the argument against utilitarianism, and relevantly similar intuitions play an epistemic role in arguments aimed at other philosophical views. Call this view *the epistemic view of intuition*. This is the view that intuitions can help overturn philosophical theories, not that intuitions are the only route by which such theories might be overturned. Many theorists defend the epistemic view of intuition overtly, offering myriad and competing views about how intuitions might play the relevant epistemic role in arguments such as the argument against utilitarianism.[7]

Both critics and advocates of the epistemic view of intuition tend to implicitly endorse an additional claim, namely that generating intuitions is the *sole* epistemic function thought experiments play in overturning philosophical theories. Thus, we find the most prominent critics of thought experiments focusing their attacks on intuitions, for instance, by claiming that intuitions: do not exist, play little or no role in extant philosophical arguments, or are systematically unreliable.[8]

Throughout most of this paper, I accept that intuitions play an epistemic role in explaining how thought experiments can help overturn philosophical theories. But I will argue that intuitions are only *part* of the story of how thought experiments can help overturn philosophical theories. Thought experiments also help clarify relevant theories, and it is only in concert with theory clarification that intuitions about cases can help overturn philosophical theories. In the relevant sense, a thought experiment *clarifies a theory* for some reasoner when that thought experiment deepens a reasoner's understanding of a theory, for instance, by improving a reasoner's grip on what the theory's concrete predictions are and how those predictions are generated by elements of the theory.

Here is the plan for the paper: I will argue that, in some cases, theory clarification can help justify the rejection of the theory clarified (§1). I will further claim that thought experiments help clarify theories without the help of intuition, even when those same thought experiments also elicit powerful intuitions (§2). Drawing on these points, I argue that thought experiments do not overturn philosophical theories by intuition alone. They do so in conjunction with theory clarification (§3). I close by sketching how a more radical view might be true, on which thought experiments help justify the rejection of philosophical theories exclusively by clarifying theories, not by any intuitions those thought experiments might generate (§4).[9]

---

raised by the argument against utilitarianism ought to be balanced against more general considerations. See McGrath & Kelly (2015) and McGrath (2019b) for the claim that levels of generality are irrelevant to philosophical theorizing.

[7] See Bengson (2014, 2015), Brogaard (2014), Chudnoff (2013, 2017), Devitt (2015), and Neta (2012).

[8] See Alexander & Weinberg (2014), Cappelen (2012), Deutsch (2015), and Williamson (2007).

[9] In contrast to the view of philosophical thought experiment I sketch in the paper's conclusion, I find it implausible that scientific thought experiments *only* function to clarify theories. Instead, I find Clatterbuck's (2013) suggestion that scientific thought experiments can justify simulation-scaffolded inductive judgments to

If I am correct that intuitions about thought experiments only overturn philosophical theories in conjunction with theory clarification, then the mainstream story about how thought experiments work is incomplete. It leaves out a critical epistemic role, one not played by intuition. As a result, the recent literature about thought experiments is misguided on both sides. On the one side, critics who decry the method of cases on the grounds that intuitions are epistemically inert or inapt have not yet given us a reason to think that thought experiments are epistemically inert. For, these arguments overlook the epistemically significant role of theory clarification. On the other side, advocates of thought experiments who have focused wholly on the epistemic value of intuition have neglected an additional route by which cases might play a role in overturning philosophical theories, namely that of theory clarification.

1 Two Routes to Rationally Discarding a Theory: Disconfirmation and Clarification

Disconfirmation and *theory clarification* are two (non-exhaustive) routes by which theorists might come to rationally discard their theories. First, a theorist might come to discard some theory by way of evidence which disconfirms a theory's first-order predictions. This is disconfirmation. For instance, suppose a theorist—let's call them Xander—provisionally accepts the theory *ants do not use tools*. Xander then discovers empirical evidence that ants sometimes stack stones to prevent competition with other colonies.[10] So, Xander rationally discards the theory in light of evidence that is first-order relative to the domain of that theory. That is, Xander discards the theory because of something they learn *about ants*. And in general, disconfirmation requires evidence about objects in the domain of the theory.

A second route by which a reasoner might come to rationally discard a theory, one which does not require new evidence about the objects in the theory's domain, is theory clarification. It occurs when a reasoner gains an improved understanding of a theory by coming to appreciate certain of the theory's predictions and how those predictions are generated by elements of the theory. This improved understanding can occur even for the reasoner who already knows what the theory says.[11]

---

be a compelling one. Here, as elsewhere, I suspect that an overly close comparison with the natural sciences has misled philosophers about the project of their own discipline (Helton forthcoming).

[10] Möglich & Alpert (1979).

[11] For present purposes, I am neutral on whether understanding is a kind of cognitive map (Gopnik, Glymour, & Sobel 2002; Gopnik et al. 2004; Grimm 2016), a skill (Hills 2016; Khalifa 2017; Cf. Sullivan 2018), a kind of relevance matching (Roush 2016, 2017), or a propositional attitude. I am employing 'understanding' in a factive way (Cf. McSweeney 2023) and am presuming that understanding is usable for the reasoner (see Elgin 2009 and Roush 2017: 404). For helpful recent overviews, see: Baumberger, Breisbart, and Brun (2017), Grimm (2021), and Hannon (2021). My notion of clarifying a theory bears important comparisons with Henk de Regt's

Consider a different theorist, call her Maya, who provisionally accepts the theory *birds do not use tools*. Maya explicitly adopts a view of tool use that is relatively minimal, as having to do with a creature's skillfully employing some object outside of its body towards some end. Maya simultaneously implicitly conceives of tool use in terms of certain anthropocentric *prototypes*, or exemplars, such as swinging a hammer and pouring liquid from a pot, and other activities involving the grasping of objects manipulated near the tool user's body.[12] As a result of Maya's reliance on anthropocentric paradigms of tool use, she fails to consider that *prey dropping* might be a form of tool use, where prey dropping is an activity employed by some birds wherein they drop prey from extreme heights to kill them. Prey dropping involves learning, skill, and repeated attempts.[13] On the minimal conception of tool use Maya endorses, prey dropping is a kind of tool use, one in which the ground itself is employed as a tool. This is so even though prey dropping involves a non-graspable object exploited far from the tool user's body.

Maya might fail to appreciate that prey dropping is, by her own lights, a form of tool use, *even if she knows everything there is to know, empirically speaking, about prey dropping*. At the same time, once Maya deepens her understanding of her theory of tool use in the relevant way, she can rationally discard her view that birds don't use tools. She can do this by relying on this improved understanding of the theory combined with the evidence she already possesses about prey dropping.[14]

At this point, one might object that theory clarification plays merely a psychological role in theory rejection, not an epistemic role. One way of developing this line of thought appeals to the familiar distinction between *propositional justification* and *doxastic justification*. A reasoner is propositionally justified in believing some claim when she has good evidence for the claim. She is doxastically justified in believing some claim when she believes that claim on good evidence. One can be propositionally justified without being doxastically justified. For instance, if Yoli has good evidence that Simon is generally a reliable person and that Simon said he'd be home by now, then Yoli is propositionally justified in believing that Simon is home by now. But if Yoli hasn't considered the matter and thus, has formed no belief concerning whether Simon is home, Yoli isn't doxastically justified with respect to this claim.[15]

---

notion of making a theory *intelligible* (de Regt 2017, de Regt & Dieks 2005) and with Michaela McSweeney's (2020) notion of a theory's being *higher-order virtuous*.

[12] Prototypes for concepts aren't typically thought to determine the extension of those concepts, so it's plausible that the exemplars Maya uses might dissociate from the extension of her concept. See Hampton (2006) and Rosch (2011) and Del Pinal (2016) for a recent discussion.

[13] Beck (1992) and Boire, Nicolakakis, & Lefebvre (2002). See Hunt, Gray, and Taylor (2013) for tool use in general.

[14] Conceptual engineering is plausibly an additional route by which theories can be rationally revised. See, e.g., Chalmers (2020), Eklund (2021), Haslanger (2020), Isaac, Koch, & Nefdt (2022), and Plunkett (2015).

[15] Silva and Oliveira (forthcoming).

Along these lines, one might suggest that Maya is propositionally justified in thinking prey dropping is a form of tool use even before she has considered the matter and thus, even before she is doxastically justified in this view. If this is right, then perhaps Maya's deepened understanding of the theory plays merely a psychological role, and not an epistemic one, in her rejection of the theory that birds don't use tools. For, perhaps Maya was already justified in rejecting the theory, even before she considered the matter. It's just that considering the matter made it psychologically possible for her to reject the theory.

This objection reflects fraught and large questions about the nature of justification, and it's not one I can do full justice to here. But to sketch a kind of reply: There is reason to think that Maya isn't even propositionally justified in rejecting her theory until she deepens her understanding of the theory in the relevant way. To see this, consider a contrast with Yoli. Yoli need only consider whether Simon is home to form the judgment that he is. The simplicity of Yoli's reasoning contributes to the sense that Yoli was already propositionally justified in thinking Simon is home by now, even before she had considered the matter.

In contrast, Maya will likely need to engage in a sustained process of reasoning to reach the judgment that prey dropping is a form of tool use. Specifically, she will have to check each element of her theory of tool use to see whether each is present in prey dropping. That is, she will have to check whether prey dropping involves (i) a skill (ii) employed towards an end (iii) partly carried out through some extra-bodily object. The fact that Maya must engage in substantial reasoning is at least weak evidence that she wasn't justified, before doing this reasoning, in thinking that prey dropping is tool use. For, the way in which Maya becomes justified in thinking prey dropping is tool use plausibly *proceeds via* Maya's insights about how the elements of her theory map onto the example of prey dropping. If this is right, Maya's deepened understanding of her theory of tool use plays an epistemic role, and not merely a psychological one, in her rejection of the theory that birds don't use tools.

I conclude that in at least some cases, theory clarification can epistemically contribute to the rational rejection of the theory clarified. In the next section, I will argue that thought experiments can clarify philosophical theories and that they can do this without the aid of intuition. Sometimes these thought experiments also elicit intuitions, but the way in which the thought experiments clarify theories is independent of those intuitions. More ultimately, I will argue that when thought experiments help overturn philosophical theories, they do so via intuition and theory clarification working together. So, for instance, the transplant case can help justify a philosopher's rejection of utilitarianism partly because it clarifies for that philosopher *what utilitarianism says about the case and why it says it*, and not (merely) by eliciting the intuition that non-consensual organ harvesting is wrong.

2 Theory Clarification from Thought Experiment, Without Intuition

The claim of this section is that reasoners can deepen their understanding of a theory by reflecting on relevant cases. Moreover, in at least some cases, this theory clarification doesn't proceed by way of intuition.[16] To see this, consider the following thought experiment:

> APPLE CASE
> Siobhan has three apples and can either give all three to Jeremiah or else can give one to Jeremiah, one to Chandra, and one to Monique. Distributing the apples in the latter way would result in more well-being than giving them all to Jeremiah, due to the diminishing marginal returns of any one person's receiving more than one apple. According to act utilitarianism, what should Siobhan do?

This case is a kind of thought experiment, but it is not one whose function is to challenge utilitarianism. Utilitarianism recommends that Siobhan distribute the apples evenly, and this is not a counter-intuitive result; indeed, reasoners might nave no intuition whatsoever about whether this is the right thing to do. At the same time, this thought experiment clarifies utilitarianism, in the sense that it can deepen a reasoner's appreciation of the what the view predicts about the case and how the elements of the theory are tied to that prediction. So, even if thought experiments do have an epistemic function of generating intuitions, this is not their only function. They can also help clarify theories, and this latter function is in principle separable from their intuition-generating function. For, in at least some cases, such as in the apple case, these functions come apart.

Here is another reason to think that thought experiments can clarify theories without the aid of intuition: Consider those minority of reasoners who, when considering the transplant case, lack a strong intuition about whether the doctor's action is morally permissible. For such reasoners, reflecting on this case might nevertheless clarify utilitarianism, insofar as it can help such reasoners appreciate what the view predicts in the case and how the elements of the view generate this prediction. So, here too we have some evidence that thought experiments can clarify theories without the help of intuition.

So far, I've argued that thought experiments can clarify theories without the aid of accompanying intuitions. But one might object that theory clarification *itself* proceeds via

[16] By arguing that thought experiments permit theory clarification without intuition, I do not mean to deny that *imagination*—both of the case itself and of how the theory might be true—might play a substantial role in theory clarification. Specifically, I take my claim to be consistent with McSweeney's (2023) important and rich development of the view that philosophy (especially metaphysics) aims at helping us to grasp a theory by imagining how the theory might be true. Intuition is a kind of rapid and non-inferential assessment, not a form of imagining. My claim is that cases permit theory clarification through a non-intuited inference, but this is consistent with this inference's itself requiring a kind of imaginative rendering of those cases and also with the resultant form of theory clarification itself being a kind of *subjectively seeming to be true* of the kind McSweeney champions as an aim of philosophy. See also McSweeney (2016) for the view that theories should be individuated at least partly by occupiable perspectives.

intuition. For instance, perhaps one just 'sees' that utilitarianism requires that Siobhan distribute the apples equally, in the sense that when considering the question, one generates a non-inferential assessment accompanied by a certain 'feel.' Or, to extend this thought to the transplant case: Perhaps reasoners just 'see' via an intuition that utilitarianism predicts that the doctor should kill the patient. If this is right, theory clarification from cases is either itself an intuition or generated by one.

Since intuitions are, at a minimum, non-inferential mental states accompanied by a 'feel,' it will suffice to show that the relevant forms of clarification do not proceed via intuition that they proceed via substantial inference (regardless of whether they also have a distinctive 'feel'). Along these lines, I will argue that the relevant assessments are inferential, even if those inferences are rapid and implicit. My argument will focus on the transplant case, but these points can be extended *mutatis mutandis* to the apple case.

Let's reflect: Does one consider the transplant case and just 'see' that the doctor's action maximizes overall well-being and hence, is what utilitarianism requires? For some reasoners, this assessment might be made extremely rapidly, such that to them it might seem to them that they simply 'see' the point, without inference. Nevertheless, there are reasons to think this assessment is in fact inferential even if, for some reasoners, this inference occurs quickly.

To make this point, I will draw on anecdotal evidence about some of the strategies commonly employed by ethics instructors who teach this thought experiment. Many of us who have taught this thought experiment will be familiar with the phenomenon that not all students immediately appreciate that the doctor's action maximizes well-being. To make this point, many of us will have attempted to motivate the point by, for instance, drawing on a chalkboard 'units' of well-being, represented, for instance, by squares—we might have resorted to jokingly calling these units of well-being 'utiles.' We might have lined up squares representing the units of well-being of the five patients whose lives were saved against squares representing the units of well-being of the patient whose life was taken, in a kind of macabre exercise of addition. This activity looks like something designed to scaffold inference. It helps students undertake a certain reasoning process, so that they can reach the conclusion that the doctor's action maximizes overall well-being.[17]

Moreover, many of us will have encountered from some of our students the objection that the doctor's action does not maximize well-being because her behavior dramatically threatens the psychic well-being of the public, who will no doubt be alarmed to learn that their local doctors are sometimes in the business of non-consensually carving up some of their patients. This objection is a good one, in that the case is not a counterexample to utilitarianism unless one can refute it. Many of us in our role as instructors will have at this point

[17] Alternatively, these activities might partly *constitute* reasoning, as defended by Dutilh Noaves (2013). Either way, the activity involves inference. Thanks to Scott Stapleford for discussion on this point.

refined the case to stipulate that the doctor's action is secret and one-off. It is not, we will stress, a medical policy, one the public is eventually bound to discover. As with the adding up of 'utiles,' this activity looks like something designed to scaffold inference. For the sake of making the strongest case possible against utilitarianism, we are attempting to lead the students to the conclusion that the doctor's action maximizes well-being and thus, is what utilitarianism requires.

One might object to these points that intuitions can be guided in certain verbally cued, inference-like ways without themselves being inferred. Elijah Chudnoff (2017) develops this general point that intuitions can sometimes be guided by verbal cues, some of which might seem to scaffold inference. For instance, he suggests that one can help a viewer 'see' both figures in this bistable stimulus in Figure 1 via instruction. For instance, one might say 'the old woman's nose is the young woman's jaw line and the old woman's mouth is a necklace on the young woman,' and in this way help a viewer see that the figure can be seen either as an old woman or as a young woman (Chudnoff 2017: 381). However, as Chudnoff argues, this guidance doesn't mean that the seeing *itself* is inferential. It is merely that these verbal guides help to alter the viewer's attention in a way that triggers the relevant visual experience.

Figure 1. A stimulus viewable as an old woman or a young woman.

Adopting this suggestion for the case of the transplant case (and extending it beyond Chudnoff's intended usage), one might suggest that the kinds of considerations ethics instructors adduce in the classroom—for instance considerations from numbers of 'utiles'—function in a similar way. These considerations help students 'see' that the doctor's action maximizes overall well-being, but they do so by triggering an intuition, not by scaffolding inference.

I cannot rule out the bare possibility that these kinds of activities—viz., those employed by ethics instructors to help students appreciate that the doctor's action maximizes well-being—function merely to trigger an intuition that the doctor's action maximizes well-being, and not to scaffold inference. However, the best explanation of these practices is that they scaffold inference, not that they merely trigger an intuition via some non-inferential route. For, consider that the claims ethics instructors adduce together entail the conclusion that the doctor's action maximizes well-being. So, it would be very surprising if the fact that

these claims can together help justify the relevant conclusion were somehow incidental to the fact that they help students form the relevant assessment.

Moreover, the ethics instructor's promptings are plausibly unlike the kinds of verbal guides one might use to try to help someone see both renderings of a bistable stimulus, such as that in Figure 1. For, the prompts employed to help one 'see' a particular rendering of a bistable stimulus likely function by directing one's visual attention in such a way so as to trigger the other visual rendering of the stimulus.[18] These promptings needn't constitute premises in an inference whose conclusion is (say) 'this is a young woman.'

I conclude that thought experiments can clarify theories and that their ability to do so does not proceed via intuition. First, the co-presence of certain common intuitions is not required for theory clarification. For instance, a reasoner might appreciate that in the apple case, utilitarianism requires that Siobhan distribute the apples evenly, without having any intuition whatsoever about whether this action is morally wrong or right. Second, the assessments by which thought experiments clarify theories are not themselves intuitions; they are inferences. For instance, the assessment that in the transplant case, utilitarianism requires that the doctor kill the patient is not itself an intuition. Rather, this assessment is an inference about an element of the case, one derived by mapping the core element of utilitarianism, the one to do with maximizing well-being, onto the case.

In the next section, I will argue that thought experiments can sometimes help overturn philosophical theories partly because those thought experiments clarify the relevant theories. At the same time, theory clarification does not play this role alone; theory clarification plays this role in concert with relevant intuitions.

3 Theory Clarification Can Help Overturn Philosophical Theories

As mentioned in the paper's outset, most theorists implicitly presume that if thought experiments help overturn philosophical theories, they do through generating intuitions. As I will now argue, this picture is at best incomplete. When thought experiments help overturn philosophical theories, they do so at least partly through theory clarification.

Consider once more the novice philosopher who: learns what utilitarianism is for the first time, provisionally accepts the view and then, after some line of reasoning which involves the transplant thought experiment, comes to reject the view. Here are three (potentially non-exhaustive) routes by which this philosopher's rejection of utilitarianism might be justified by the thought experiment:

DISCONFIRMATION ALONE

[18] Tsal & Kolbet (1985) and Long & Toppino (2004).

> Via the thought experiment, the philosopher comes to appreciate something **about morality**, and this fact *alone* explains the fact that the thought experiment helps to justify her rejection of utilitarianism.
>
> THEORY CLARIFICATION ALONE
> Via the thought experiment, the philosopher comes to appreciate something **about utilitarianism**, and this fact *alone* explains the fact that the thought experiment helps to justify her rejection of utilitarianism.
>
> BOTH DISCONFIRMATION & THEORY CLARIFICATION
> Via the thought experiment, the philosopher comes to appreciate **both something about morality and something about utilitarianism** and these facts together explain the fact that the thought experiment helps to justify her rejection of utilitarianism.

I will argue against the 'disconfirmation alone' explanation. The fact that the transplant case helps justify the philosopher's rejection of utilitarianism is not explained wholly by the fact that the case helps her appreciate something about morality. Rather, this explanation must appeal, at least in part, to the fact that the case helps the philosopher appreciate something about utilitarianism. Specifically, I will develop the view that the case helps the philosopher appreciate both something about morality and something about utilitarianism and that these epistemic improvements *together* help justify her rejection of utilitarianism.

Here is how the transplant case helps to justify the philosopher's rejection of utilitarianism. First, the case helps to generate the philosopher's intuition that the doctor's killing of the patient is morally impermissible. Second, the case helps to elicit the philosopher's inference that utilitarianism requires that the doctor kill her patient. Notice that while the first assessment is intuitive and the second is inferred, both are about the very same action, the doctor's killing of her patient. Moreover, it is essential for the thought experiment to do its work that it be the very same action of the doctor which is both morally impermissible and mandated by utilitarianism. For, it is only because the case justifies the reasoner in thinking that the very same action has both of these traits simultaneously that the case constitutes a counterexample to utilitarianism. To see this, consider once more the argument against utilitarianism:

> **Argument Against Act Utilitarianism**
> (1) The doctor's action in the transplant case is morally wrong.
> (2) The doctor's action (the same action as in (1)) maximizes well-being.
> (3) According to act utilitarianism, the doctor's action in the transplant case is not morally wrong.

(4) Act utilitarianism is false.

In this formulation of the argument, (1) and (2) must be about the same action. Otherwise, the argument would not be valid. While this point is trivial enough, it is not trivial that reasoners can become justified in holding (1) and (2) simultaneously, through reflection on the transplant case.[19] When this happens, it is because of the intuitive justification of (1) and the inferred justification of (2). In other words, it is through disconfirmation and theory clarification, working together, that the case somehow helps to justify the argument. So, it is false that the case buttresses the argument solely by enhancing the philosopher's appreciation of moral facts. Of at least equal importance, the case permits a kind of inference about utilitarianism, and this latter inference is also critical in the philosopher's being justified in rejecting utilitarianism.

What is it about the thought experiment that explains how it can justify a reasoner in thinking that the same action is simultaneously morally impermissible and mandated by utilitarianism? This question is entangled with multiple empirical and epistemological questions, and it is not one I could hope to answer here. But, as a first attempt at an answer: Plausibly it is something about the specificity of the thought experiment that justifies this assessment, much as some models, in virtue of their specificity, permit reasoners to see how multiple traits can inhere simultaneously in the same entity.

Compare: One might point to a topographical map of the Earth to come to see how a location could both be ~4,000 feet in elevation and located in the Northern Hemisphere. The Mojave Desert is such a location and reading this kind of map can justify a reasoner in thinking that the Mojave has both such traits. This isn't to suggest that a topographical map is the only way to be so justified; testimony could also play this role. Rather, the present claim is that when it is the map which justifies one in thinking that the Mojave has both these traits, there is something about the specificity of the model which explains how it justifies this. Likewise, something about considering a particular action of the doctor helps reasoners to appreciate that the *very same* action which is morally impermissible is also mandated by utilitarianism.[20] In its specificity, the thought experiment facilitates seamlessly coordinated reasoning about features simultaneously had by the same action.[21]

---

[19] I do not claim that it is *only* through thought experiment that one might be simultaneously justified in holding (1) and (2). For instance, perhaps testimony could confer this kind of justification. See McGrath (2019a: 59-105; 2021) for a defense.

[20] Philosophical use of thought experiments is perhaps comparable to the use of concrete examples in math and science learning, which can help facilitate more abstract forms of reasoning. See Fife et al. (2014), Goldstone & Son (2005), Koedinger et al. (2008), Novak et al. (2014), and Schwartz et al. (2011).

[21] To suggest that thought experiments have a kind of specificity is not to suggest that they are tokens as opposed to types. Even if thought experiments are types, they still might permit us to focus on specific aspects of that type. Special thanks to Elijah Chudnoff for discussion on this point.

I conclude that the fact that thought experiments can help to overturn philosophical theories is not wholly explained by the intuitions that those thought experiments generate. Rather, thought experiments help to overturn philosophical theories at least partly through helping to clarify the very theories which are overturned.

Think once more of Maya, the theorist who provisionally accepted that birds do not use tools, even though she had evidence that birds engage in prey dropping. It was only after coming to appreciate that her own theory counted prey dropping as tool use that Maya was justified in revising her view. In other words, Maya's empirical evidence about the objects in the theory's domain was inadequate to justify the rejection of her view until she deepened her understanding of the theory. Likewise, for the philosopher who contemplates the organ harvesting case, it is partly because the case reveals that sometimes non-consensual organ harvesting *is* a way of maximizing well-being and thus is required by utilitarianism that the case can help justify the philosopher's rejection of utilitarianism.

4 Is Intuition a Red Herring?

I have argued that thought experiments can facilitate both disconfirmation of a theory and clarification of the same theory and that it is only when these functions work together that thought experiments can help overturn philosophical theories. For instance, the transplant thought experiment helps to facilitate an *intuition* that the doctor's killing of the patient is morally impermissible, and it also helps to ground an *inference* that utilitarianism morally requires that the doctor kill the patient. And both this intuition and this inference, made about one and the same action, are essential in explaining how the case can help to justify the rejection of utilitarianism.

These claims have important implications for metaphilosophy. First, as previously mentioned, some theorists criticize the role of thought experiments in philosophy by way of criticizing the epistemic role of intuitions. These theorists implicitly presume that if thought experiments play any epistemic role in theory revision, it is exclusively by way of intuitions, such that criticizing the epistemic status of intuitions amounts to a criticism of the method of cases. But if the points of this paper are correct, even if critics of the method of cases are right that intuitions never epistemically matter in overturning philosophical theories, it wouldn't follow that thought experiments are epistemically inert in overturning such theories. To make this further claim, they must show that thought experiments play no epistemic role in *theory clarification* or else that theory clarification is epistemically inert in the overturning of philosophical theories. By the same token, advocates of the method of cases, in focusing almost exclusively on the epistemic role of intuition, have neglected a distinct way in which cases can help justify the rejection of philosophical theories.

I close with a further speculative and more radical thought. I have been supposing that the reason thought experiments can help overturn philosophical theories is partly

because of the intuitions which such cases can elicit. For instance, I have been supposing that the intuition *it is wrong for the doctor to kill her patient* partly explains how the transplant case justifies a rejection of utilitarianism. But now, I would like to briefly entertain a rival view on which intuitions about cases play no epistemic role in the overturning of philosophical theories.

Here is one way intuitions might be epistemically inert in the overturning of philosophical theories. Suppose that the typical reasoner who considers the transplant case is *already justified*, before reflecting on the case, in thinking that non-consensual transplant is morally wrong. In a certain light, this supposition isn't just coherent; it is nearly undeniable. For how could a typical adult fail to possess good evidence that it's wrong to carve someone up without their consent? If this is right, then this reasoner's intuition about the wrongness of the doctor's action does not obviously play an *epistemic* role in her rejection of utilitarianism, even though it does play a *psychological* role in explaining her rejection. For, the intuition she has about the case does not constitute new evidence for her. It merely encodes or reflects evidence she already possesses. At the same time, this intuition plausibly is psychologically necessary in her rejection of the theory, since she presumably needs to token the relevant judgment in order to be able to use it in the service of an argument against utilitarianism.

Suppose further that the core points of this paper concerning theory clarification are correct: Thought experiments can play the epistemic role of helping philosophers to clarify theories; theory clarification in turn plays an epistemic role in helping to justify the rejection of certain philosophical arguments; and theory clarification is not driven by intuition. Taken together, these points would suggest that the mainstream picture of thought experiments as functioning primarily in terms of intuitions is not just incomplete. It is entirely misguided. For, on this view, thought experiments play an epistemic role in overturning philosophical theories, but not through whatever intuitions they might generate.

On this more radical view I have tentatively sketched, we should think of cases such as the transplant case as primarily functioning to tell us something about utilitarianism, the ethical *theory*, not about morality itself. Likewise, we should take cases such as the Gettier case as primarily functioning to tell us something about the tripartite *theory* of knowledge, not about knowledge itself. I leave the viability of such a view as a further question.[22]

WORKS CITED

[22] For extremely helpful feedback on this paper, I am indebted to: Josh Armstrong, Elijah Chudnoff, Kevin McCain, Michaela McSweeney, Chris Register, Scott Stapleford, and audience members at the Workshop on Seemings and Intuitions. Special thanks to Mark van Roojen for inspiring this paper by way of a remark, made offhand in an undergraduate ethics course of his in which I was a student, that in doing ethics, we use moral knowledge we already have.

Alexander, J. & Weinberg, J. M. (2014). The "unreliability" of epistemic intuitions. In *Current controversies in experimental philosophy* (pp. 128-148). Routledge.

Beck, B. B. (1982). Chimpocentrism: Bias in cognitive ethology. *Journal of Human Evolution*, *11*(1), 3-17

Bengson, J. (2014). How philosophers use intuition and 'intuition.' *Philosophical Studies*, *171*(3), 555-576.

Bengson, J. (2015). The intellectual given. *Mind*, *124*(495), 707-760.

Boire, D., Nicolakakis, N., & Lefebvre, L. (2002). Tools and brains in birds. *Behaviour*, *139*(7), 939-973.

Brogaard, B. (2014). Intuitions as intellectual seemings. *Analytic Philosophy*, *55*(4)

Brown, J. R. and Yiftach F.,(2022). Thought Experiments, *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2022/entries/thought-experiment/>.

Cappelen, H. (2012). *Philosophy without intuitions*. Oxford University Press.

Chalmers, D. J. (2020). What is conceptual engineering and what should it be? *Inquiry*, 1-18.

Chudnoff, E. (2011). What intuitions are like. *Philosophy and Phenomenological Research*, *82*(3), 625-654.

Chudnoff, E. (2013). *Intuition*. Oup Oxford.

Chudnoff, E. (2017). The reality of the intuitive. *Inquiry*, *60*(4), 371-385.

Clatterbuck, H. (2013). The epistemology of thought experiments: A non-eliminativist, non-platonic account. *European Journal for Philosophy of Science*, *3*(3), 309-329.

Del Pinal, G. (2016). Prototypes as compositional components of concepts. *Synthese*, *193*, 2899-2927.

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.

de Regt, H. W. & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 137-170.

Deutsch, M. E. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method*. Mit Press.

Devitt, M. (2015). Relying on intuitions: Where Cappelen and Deutsch go wrong. *Inquiry*, *58*(7-8), 669-699.

Dutilh Noaves, C. D. (2013). Mathematical reasoning and external symbolic systems. *Logique et analyse*, *56*(221), 45-65.

Elgin, C. (2009). Is understanding factive? *Epistemic value*, 322-330.

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational psychology review*, *26*, 9-25.

Goldstone, R. L. & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the learning sciences*, *14*(1), 69-110.

Gopnik, A., Glymour, C., & Sobel, D. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. *The cognitive basis of science*. 117-132.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. Psychological review, 111 (1), 3.

Gopnik, A. & Schwitzgebel, E. (1998). Whose concepts are they, anyway? The role of philosophical intuition in empirical psychology. *Rethinking intuition*, 75-91.

Grimm, S. (2016). Understanding and transparency. In Explaining Understanding (pp. 228-245). Routledge.

Grimm, S. (2021). Understanding. The Stanford Encyclopedia of Philosophy , Edward N. Zalta
(ed.), https://plato.stanford.edu/archives/sum2021/entries/understanding/

Hampton, J. (2006). Concepts as prototypes. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 79–113). Amsterdam: Elsevier.

Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly*, *58*(3), 269-290.

Haslanger, S. (2020). How not to change the subject. *Shifting Concepts: The Philosophy and Psychology of Conceptual Variation*, 235-259.

Helton, G. (forthcoming). Viewpoint convergence as a philosophical defect, *Attitude in Philosophy*, eds. Goldberg & Walker, Oxford University Press.

Hunt, G. R., Gray, R. D., & Taylor, A. H. (2013). Why is tool use rare in animals. *Tool use in animals: cognition and ecology*, 89-118.

Ichikawa, J. & Jarvis, B. (2009). Thought-experiment intuitions and truth in fiction. *Philosophical Studies*, *142*(2), 221-246.

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*, *17*(10), e12879.

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, *32*(2), 366-397.

Koksvik, O. (2011). *Intuition*, Ph.D. Thesis, Australian National University

Long, G. M. & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: alternating views of reversible figures. *Psychological bulletin*, *130*(5), 748.

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, Philosophy and the Empirical*, *31*(1), 128-159.

Malmgren, A. S. (2011). Rationalism and the content of intuitive judgements. *Mind*, *120*(478), 263-327.

McGrath, S. (2019a). *Moral knowledge*. Oxford University Press.

McGrath, S. (2019b). Philosophical methodology and levels of generality. *Philosophical Perspectives*, *33*(1), 105-125.

McGrath, S. (2021). Epistemic autonomy for social epistemologists: The case of moral inheritance. In *Epistemic Autonomy* (pp. 271-287). Routledge.

McGrath, S. & Kelly, T. (2015). Soames and Moore on method in ethics and epistemology. *Philosophical Studies*, *172*, 1661-1670.

McSweeney, M. M. (2016). An epistemic account of metaphysical equivalence. *Philosophical Perspectives*, *30*, 270-293.

McSweeney, M. M. (2020). Theories as recipes: third-order virtue and vice. *Philosophical Studies*, *177*(2), 391-411.

McSweeney, M. M. (2023). Metaphysics as Essentially Imaginative and Aiming at Understanding. *American Philosophical Quarterly*, *60*(1), 83-97.

Möglich, M. H. & Alpert, G. D. (1979). Stone dropping by Conomyrma bicolor (Hymenoptera: Formicidae): a new technique of interference competition. *Behavioral Ecology and sociobiology*, 105-113.

Neta, R. (2012). Knowing from the armchair that our intuitions are reliable. *The Monist*, *95*(2), 329-351.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction: Using the hands to learn math. *Psychological science*, *25*(4), 903-910.

Plunkett, D. (2015). Which concepts should we use?: Metalinguistic negotiations and the methodology of philosophy. *Inquiry*, *58*(7-8), 828-874.

Pust, Joel, (2000). *Intuitions as evidence*, New York: Garland/Routledge.

Pust, Joel, (2019). Intuition. *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/intuition/>.

Rosch, E. H. (2011). Slow lettuce: Categories, concepts, fuzzy sets, and logical deduction. In R. Belohlavek & G. J. Klir (Eds.), *Concepts and fuzzy logic* (Chap. 4). Cambridge, MA: The MIT Press.

Roush, S. (2016). Simulation and Understanding Other Minds. *Philosophical Issues*, *26*(1), 351-373.

Roush, S. (2017). The difference between knowledge and understanding. *Explaining Knowledge: New Essays on the Gettier Problem*, 384.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of educational psychology*, *103*(4), 759.

Silva, P. & Oliveira, L. R. G. (forthcoming). Propositional justification and doxastic justification. In M. Lasonen-Aarnio & C. M. Littlejohn (Eds.), *Routledge handbook of the philosophy evidence*. Routledge.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 204-217.

Tsal, Y. & Kolbet, L. (1985). Disambiguating ambiguous figures by selective attention. *The Quarterly Journal of Experimental Psychology Section A*, *37*(1), 25-37.

Williamson, T. (2007). *The Philosophy of Philosophy*, New York: Routledge.