To appear in Michael T. Stuart; Yiftach Fehige; James Robert Brown (eds). 2017. *The Routledge Companion to Thought Experiments*. Abingdon/New York: Routledge. 195–210.

Thought Experiments in Ethics

Georg Brun
Georg.Brun@philo.unibe.ch

Abstract This chapter suggests a scheme of reconstruction, which explains how scenarios, questions and arguments figure in thought experiments. It then develops a typology of ethical thought experiments according to their function, which can be epistemic, illustrative, rhetorical, heuristic or theory-internal. Epistemic functions of supporting or refuting ethical claims rely on metaethical assumptions, for example, an epistemological background of reflective equilibrium. In this context, thought experiments may involve intuitive as well as explicitly argued judgements; they can be used to generate moral commitments, to explore consequences of moral theories, and to show inconsistencies within or between moral commitments and moral theory; but the results of thought experiments by themselves do not settle what is epistemically justified and may also be rejected. Finally, some prominent challenges are discussed: do unrealistic scenarios undermine epistemic thought experiments? Are ethical thought experiments misleading? Do they rely on weak analogies? Are there specifically moral objections to ethical thought experiments?

1. Introduction

In normative and applied ethics, thought experiments are frequently used in debates on questions such as: Are we morally obliged to give large amounts to poverty relief? Can the number of lives saved be ethically decisive? Is torture ever morally permissible? At the heart of every thought experiment is a scenario with a question. Here are sketches of some of the best-known ethical thought experiments, which have also become prominent in political discourse:

Trolley: in the so-called *Driver's Two Options* (there are many other versions), a runaway trolley is heading towards five workers on the track who have no escape; is it morally permissible for the driver to steer the trolley onto another track where it will kill one worker? (Foot 2002; Thomson 2008)

Pond: a child has fallen into a pond and is about to drown; ought you hurry to rescue him even if doing so ruins your shoes and thwarts your plans for the day? (Singer 1972)

Violinist: while asleep, your body has been hooked up to the body of a violinist who has an illness that makes it necessary for her to be connected to your metabolism for nine months; is it morally permissible for you to ask to be severed from her even if this will kill her? (Thomson 1972)

Ticking Bomb: a terrorist has hidden a bomb that will kill thousands of people if set off; are we morally permitted to torture her given that we know that only in this way we will learn how to defuse the bomb? (Shue 1978)

Original Position: behind a "veil of ignorance," in a situation in which no one knows his place in society, his status or his natural strengths and weaknesses, which fundamental principles of justice would free and rational persons accept as being in their interest? (Rawls 1999)

In contrast to the first four scenarios, *Original Position* is not a stock example. One explanation for this is that *Original Position* has a specific function in Rawls's theory of justice, whereas the other scenarios are used in more familiar ways as parts of arguments for or against an ethically relevant claim about, for example, abortion, torture or our obligation to help others. However, the first four scenarios are sometimes used differently as well, for example, to illustrate the consequences of some moral principle. This calls for a typology of ethical thought experiments according to their functions, which will be introduced in Section 3. Before this issue can be tackled, we need a more detailed characterization of the structure of thought experiments which explains how scenarios, questions and arguments figure in thought experiments (Section 2). Section 4 then investigates the two most widely discussed functions of thought experiments, namely supporting and refuting ethical claims. Since such epistemic functions cannot be analysed in a metaethically neutral way, I introduce some elements of the method of reflective equilibrium as a more specific background in moral epistemology, which enjoys relatively broad acceptance. In Section 5, I finally discuss a range of challenges and touch upon specifically moral objections to ethical thought experiments.

2. Reconstructing ethical thought experiments

The philosophical literature usually focuses on analysing thought experiments of a certain kind, most often, thought experiments that aim at refuting a general claim. For such thought experiments, specific schemes of reconstruction have been developed, typically tied to some philosophical approach (e.g. Sorensen 1992, ch. 6; Häggqvist 1996, ch. 5, 2009; and see the entries in Part III of this volume). In this chapter, I rely on a simpler reconstruction, which enables us to address the variety of ethical thought experiments and to distinguish two common uses of the term “thought experiment”:

(1) A scenario and a question are introduced.

(2) The experimenter goes through (imagines, thinks about, etc.) the scenario and arrives at some result.

(3) A conclusion is drawn with respect to some target (e.g., an ethically relevant claim or distinction).

A “core” thought experiment just comprises (1) and (2). An “extended” thought experiment includes some additional reasoning (3), which can be reconstructed as an argument; the result of (2) is used as a premise from which, typically together with a range of often implicit assumptions, a conclusion is drawn with respect to some target.

In *Trolley,* for example, the scenario describes a situation with a runway trolley, two tracks, a driver, several workers and a switch. The question is whether the driver is morally permitted to throw the switch. The experimenter imagines or thinks about the situation and arrives at, say, the conviction that the driver is permitted to change the switch. This is the result of the core thought experiment. The extended thought experiment then uses this result as a premise in an argument for, say, the target thesis that (body) numbers count morally. An additional assumption involved in this argument may be that the difference in numbers of workers on the tracks is the reason why switching is permissible.

In more complex cases, one extended thought experiment draws on the results of several core thought experiments. Smart (1972, 27–8), for instance, uses two scenarios, one with one and another with two million equally happy people, to illustrate the difference between average (both scenarios yield the same result) and total utilitarianism (the situation with more people fares better).

Conversely, the same core thought experiment can be used in different arguments. This provides a more natural and flexible analysis of reasoning by thought experiment than an analysis which uses “thought experiment” only for extended thought experiments, associating every thought experiment with a specific structure (1)–(3) (see, e.g., Häggqvist 1996, 2009).

The elements (1)–(3) require several explanations. The scenario describes a situation which is declared or assumed to be possible in some sense, for example, (meta)physically, logically or conceptually.[1] If the thought experiment is to be successful, the scenario must be a consistent, meaningful description (see Rescher 2005, ch. 9) and the person running the thought experiment must be able to make appropriate use of the scenario by, for example, conceiving of the described situation or applying a given principle to it. Often, the scenario is qualified as fictional (e.g. Elgin 2014), hypothetical or counterfactual (e.g. Gendler 2000, 17). This is not meant to imply that the described situation must not be real or realizable, but only that it need not be so, and that, if the described situation happens to be realized, this does not affect the thought experiment. Passersby come across drowning children, and *Trolley*-style accidents do in fact happen, but as thought experiments, *Pond* and *Trolley* are independent of those tragedies. This is so because even if situations as described in the scenario are real, the scenario, not these situations, sets the stage for the thought experiment. As all descriptions of non-abstract states of affairs, scenarios are invariably less rich than the actual and possible situations that fit the description (does the violinist like eggplant casserole?). This selectivity serves to establish at least a presumption about which aspects of a situation are relevant for answering the given question, without implying the unrealistic assumption that the scenario explicitly mentions all and only the relevant information. Common ground and implicatures play an important role as well.

The element (2) is deliberately framed in unspecific terms to leave room for a broad range of activities and results. More specific descriptions can be used to characterize specific kinds of thought experiments, for example, thought experiments in which an intuition is prompted by imagining a situation or thought experiments which use modal reasoning. We should, however, resist the temptation to postulate any such specific description as a general characterization. It is, for example, sometimes said that ethical thought experiments are used to elicit intuitions (see Stich and Tobia, this volume). What this boils down to depends on what one means by "intuition." A necessary but not sufficient requirement for a belief (or other propositional attitude) to be an intuition is that it is not held *just* because it has been inferred consciously (see Brun 2014). But as a general characterization of the result in (2), this minimal condition is still too narrow.[2] There are many ethical thought experiments in which the result of (2) is a judgement gained by means of an explicit argument, typically involving additional assumptions not explicitly mentioned in the scenario.

Similar problems result if thought experiments are distinguished from counter-factual or hypothetical reasoning in general by the requirement of having an experimental element in the sense that something is imagined not merely by contemplating propositional content, but visually, tactually or in another perception-related way (Brown and Fehige 2014). If this idea is to be effective at all, the element of experience must be more than the visual, auditory or other associations routinely evoked by any description of a non-abstract situation. Understood in this way, however, an implausibly narrow notion of thought experiment results, which excludes many philosophical (e.g. Putnam's H_2O vs. XYZ on Twin Earth) and ethical thought experiments such as *Ticking Bomb* and *Trolley* if they are described as simply as at the beginning of this section.

The argumentation reconstructed in (3) may take any form of argument ranging from modal and deontic deductive reasoning to arguments from analogies and inference to the best explanation. The target mentioned in (3) allows for many different kinds of conclusions: they can be about a concept, a proposition, a theory, or about a relation between two or more such objects (e.g. a difference between two theories). As Gendler (2000, 25) argues, there is a tendency (but no strict rule) that thought experiments in different disciplines aim at different results in (2). Scientific thought experiments primarily ask what would happen in the situation described in the scenario; thought experiments in metaphysics, epistemology and philosophy of language more often ask how we should describe what would happen; and thought experiments in ethics and aesthetics very commonly ask how what would happen should be evaluated. Accordingly, scientific thought experiments typically target empirical claims; in metaphysics, epistemology and philosophy of language, the target is typically a proposal for a conceptual analysis or, more generally, a characterization of a concept (e.g. Gettier-cases against knowledge as justified true belief); in ethics, the target is usually a claim about a normative concept or some normative principle or theory, for example, that numbers count morally, that abortion is permissible, or that the right to life includes a right to be sustained in living.

The explanations given so far obviously do not amount to a definition of "ethical thought experiment," but together with the examples above they should suffice to give a fairly clear idea of what ethical thought experiments are. On this basis, we can now explore the functions of ethical thought experiments in more detail.

3. Functions of ethical thought experiments

Typologies of extended thought experiments according to their goal are of special importance since the evaluation of a thought experiment must frequently be relative to what it is supposed to achieve. In what follows, I suggest a typology that can accommodate the broad range of functions of thought experiments in ethics (for alternatives, see, e.g., Brown 1991; Walsh 2011; Davis 2012). It is important to keep in mind that core thought experiments are normally put to more than one use, and that their use can change over time, for instance, because they are employed in the context of different background assumptions and with different goals (Elgin 2014). The proposed classification therefore primarily deals with extended thought experiments, and only derivatively with core thought experiments.

To simplify, we can speak of, for example, an "epistemic thought experiment" where we strictly speaking would have to say "extended thought experiment with an epistemic function" or "core thought experiment used in an extended thought experiment with an epistemic function."

3.1. Epistemic thought experiments

The goal of epistemic thought experiments is to provide a reason which speaks in favour of or against a claim with respect to the target.[3] Accordingly, we can draw a basic distinction between constructive and destructive epistemic functions. Constructive thought experiments come in two varieties. Some are meant to show that something is possible. Shue (1978), for example, uses *Ticking Bomb* to show that there are possible exceptional circumstances in which torture is permissible, and according to one analysis, *Violinist* aims at showing "that abortion could be morally permissible even when the fetus has a right to life" (Brown and Fehige 2014). Other constructive thought experiments are intended to provide only some (non-conclusive) support for a claim. *Pond* is regularly interpreted in this way, either as intended to support a general principle that obliges us to help if we can do so without significant moral costs, or as an argument from analogy which supports the claim that we should give significant amounts of money to help poor people. That thought experiments can provide support for an ethical principle at all, is a view that needs metaethical backing. On Singer's own view (2009, 15–17), for example, intuitions are epistemically dubious, including those prompted by *Pond;* and Dancy (e.g., 1985, 1993, 64–6) has argued that analogies from core thought experiments to real cases are inherently problematic. Section 4 discusses epistemic functions in more detail against a specific background in moral epistemology.

Destructive thought experiments provide a counterexample to a target claim. Often, they show that something can be the case and use this in an argument against an incompatible claim. *Violinist,* for example, is standardly reconstructed as providing such a refuting argument against the claim that the right to life of one person under all circumstances outweighs the right of another person to decide what happens with his or her body. Obviously, this type of refuting use of a scenario is most closely related to the possibility-showing constructive use. *Ticking Bomb* shows that torture is permissible in some circumstances just in case it refutes the claim that torture is impermissible under all conditions whatsoever. Another way of using the result of going through the scenario is to turn it against some background assumption. *Violinist*, for example, can also be reconstructed as challenging the tacit assumption that the right to life entails a right to be sustained in living (Brown and Fehige 2014).

It is natural to think that epistemic functions of thought experiments are normally refuting rather than constructive because general claims can be decisively undermined but not conclusively supported by individual cases. And in fact, there is a culture of "counterexample philosophy," which widely employs refuting thought experiments which target a conceptual claim that is part of some account of, for example, moral permissibility, autonomy or consent. The thought experiment then comes to the result that an act described in the scenario does (not) fall under that concept in contrast to what follows from the account in question. Whether that actually amounts to a refutation of this account depends on methodological assumptions, in particular on the conditions an adequate "analysis" or "explication" of a concept is supposed to meet. If the goal is an analysis which preserves all clear-cut cases, counterexamples to such cases invariably refute the proposed analysis. If the goal is an explication that may include some conceptual revision, counterexamples are not automatically decisive.

In the philosophical discussion, epistemic or, more specifically, destructive epistemic functions are generally taken to be the basic or the most important functions of thought experiments. There is, however, not one function all ethical thought experiments essentially have. And even if one considers non-epistemic functions as entirely secondary, it remains important to recognize them because ethical thought experiments usually have an epistemic function non-exclusively, and therefore other, non-epistemic functions can influence the epistemic function or the two functions can get conflated.

3.2. Illustrative and rhetorical thought experiments

Illustrative thought experiments render a concept, a proposition, a theory or some other target clearer and more understandable, but insofar as we focus exclusively on its illustrative function, we do not see a thought experiment as providing reason for or against its target. Singer's original use of *Pond* is clearly intended as an illustration of a principle he independently defends. He writes (1972, 231): "If it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it. An application of this principle would be as follows." The case of the drowning child makes then vivid what the quoted principle may ask of us in a concrete situation. Another example is the difference between total and average utilitarianism. One may know in abstract terms that it will become manifest only in a comparison of two groups with a different number of people, but going through specific scenarios makes the difference more readily graspable. Furthermore, textbook-examples which are thought experiments are usually illustrative (Cohnitz 2006, 77).

Closely related are rhetorical thought experiments, which are intended to convince people to accept or reject some ethical idea. *Pond* is certainly apt for rhetorical use and actually functions this way in some of Singer's later writings (e.g. 2009). Rhetorical functions very often piggyback on an epistemic function, but certainly not always. Ticking-bomb scenarios, for example, are notorious for having been used in media and political discourse as a means for campaigning in favour of torture and other illegal or contested practices.

Of course, illustrative thought experiments can be misleading, and rhetorical thought experiments can be used in propagandistic or demagogic ways, but this is not necessarily so. We expect successful epistemic thought experiments to have the rhetorical power of convincing us, and there is nothing inherently problematic about that, in contrast to an illustrative or rhetorical function of a thought experiment which is interpreted or sold as epistemic (for a paradigmatic example, see Dennett's 1980 critique of Searle's Chinese room argument as an "intuition pump").

3.3. Heuristic thought experiments

In the simplest case, heuristic thought experiments are just devices for generating ideas. A core thought experiment is run, but its result is then investigated independently. More ambitious heuristic thought experiments systematically explore morally relevant factors and differences.

One possibility is to try to isolate factors which are potentially morally relevant by developing variations of a scenario and seeing whether the resulting moral intuitions or judgements co-vary. Thomson’s work on trolley cases is a well-known example. She extensively varies the scenario in order to find whether it makes a difference if, for example, the driver or a bystander can change the switch, or if the choice is described as one between killing five and killing one, or between killing one and letting five die. Actually, this procedure can also run into problems. Many more factors can be tested for moral relevance, including the age of the people on the track, whether they are employees of the train-system, or whether they are members of an ethnic minority. In fact, there is empirical research showing that skin colour has an influence on the judgement of people who are strongly committed to its moral irrelevance (see Uhlmann et al. 2009). This raises the question of whether such an effect undermines the value of the elicited judgements, or whether we should rather conclude that a trolley scenario which alludes to skin colour is badly designed, or that philosophers need to be more careful in running thought experiments (see Stich and Tobia, this volume; Ludwig, this volume).

Another heuristic strategy systematically explores the consequences of various theories for a range of scenarios with the goal of, for example, finding out in which situations two rival theories actually lead to different results. Again, comparisons of various versions of utilitarianism provide examples.

Applications of the two strategies just described can actually be analysed as heuristic or as another type of thought experiment, depending on what we take to be their target. Consider the example of the two versions of utilitarianism. If the target is the claim that there is a difference between total and average utilitarianism, then Smart’s thought experiment is most naturally interpreted as illustrating this difference or as having the epistemic function of showing that there is indeed such a difference. If the target is, say, total utilitarianism, then Smart’s thought experiment may be interpreted as having the heuristic function of tracking down a situation in which total utilitarianism differs from average utilitarianism, but in itself this neither supports nor undermines total utilitarianism. The last interpretation shows that the heuristic potential of thought experiments can be exploited even by those who are deeply sceptical about epistemic functions of ethical thought experiments (e.g. Dancy 1993, 65).

In the literature there is a tendency to see illustrative and heuristic functions of thought experiments as of merely limited interest, especially in comparison with epistemic functions. However, another view emerges, for example, in Hills’s recent discussion of moral understanding. According to her account, to understand that *q* is why *p* is morally right or wrong, requires that one has the ability to “draw the conclusion that *p*' (or that probably *p*') from the information that *q*' (where *p*' and *q*' are similar to but not identical to *p* and *q*)” (Hills 2009, 102). In other words, one needs the ability to run heuristic and illustrative thought experiments. Furthermore, Hills argues that moral understanding is not only instrumental in reliably doing the right thing and justifying yourself to others, but also essential for having a good character and for morally worthy action. On this account, the ability to run thought experiments is of central interest not only to ethical theorists, but to all moral agents.

3.4. Thought experiment with a theory-internal function

Borsboom, Mellenbergh and Van Heerden (2002) argue that there are thought experiments which perform some specific function within a theory, different from the functions discussed so far. Their prime example is the scenario of a “counterfactual long run” in frequentist statistics. A prominent example of an ethical thought experiment with a theory-internal function is Rawls's *Original Position.* The scenario models fair conditions for social cooperation as well as restrictions on arguments for principles of justice (Rawls 1999, 16–9; 2001, 17, 83, 85; cf. Gendler 2007 for a different analysis). Also Kant's categorical imperative involves a thought experiment in which we have to test whether we can will the generalization of the maxim guiding some way of acting (see Parfit 2011, 285, 328–9). A closer analysis and evaluation of such thought experiments inevitably requires an in-depth discussion of the respective ethical theories and cannot be undertaken here.

4. Ethical thought experiments and reflective equilibrium

Epistemic functions of thought experiments are philosophically the most prominent and hence deserve a closer analysis. But as pointed out in Section 3.1, such functions depend on metaethical assumptions. I therefore introduce in this section a background in moral epistemology, which is relatively widely accepted, namely reflective equilibrium (Section 4.1). On this basis, the central functions of thought experiments can be characterized more precisely (Section 4.2), and some consequences can be drawn regarding the results of core thought experiments and with respect to what we can expect to accomplish by thought experiments (Section 4.3).

4.1. Reflective equilibrium

The method of (so-called "wide") reflective equilibrium can be characterized most succinctly by two key ideas. Firstly, judgements and principles are justified if judgements, principles and background theories are in equilibrium. Secondly, this state is reached through a process that starts from judgements and background theories, proposes systematic principles and then mutually adjusts judgements and principles (and possibly also background theories).[4]

The contrast between principles and judgements is not to be understood as a matter of their content, but as the contrast between propositions that are part of some given moral theory or system of principles and propositions to which somebody is actually committed. Commitment is to be understood as an epistemically relevant status which implies at least a minimal degree of credibility. Commitments can be expressed explicitly or merely revealed in action and they can have any degree of firmness from feeble to unwavering. In the tradition of Rawls and Daniels, commitments are required to be considered judgements; that is, judgements not made under circumstances prone to error such as inattention, excessive self-interest, etc. Both, commitments and elements of the theory, may be changed in the process of mutual adjustment. Every stage of the process of developing a reflective equilibrium is characterized by a position consisting of the current commitments and the current theory.

4.2. Functions of thought experiments

Within the framework of reflective equilibrium, thought experiments can play several roles. Constructive epistemic thought experiments can generate commitments at any stage in the process of developing a reflective equilibrium. In simple cases, the commitment is just the result of the core thought experiment. The experimenter considers *Trolley,* for example, and resolves that throwing the switch is morally permissible. In more complex cases, a conclusion of an extended thought experiment is supported by the result of a core thought experiment and additional reasoning. In this way, *Trolley* may support the claim that it is morally permissible to kill a person if this is the only way to save several others. While such supportive thought experiments can be independent of the current theory, this need not be the case. Elements of the current theory can enter the reasoning, and candidates for commitments can also be generated just by applying the current theory to the scenario. In *Trolley,* a philosopher with deontologist leanings could argue that the prohibition on killing implies that it is morally impermissible to change the switch, and then conclude that this result should be accepted because it does not conflict with any of her other commitments.

Of course, conflicts will often arise, and then destructive thought experiments become important. In such cases, the result of a core thought experiment is a premise in an argument which shows that there is an incoherence within or between the current commitments and the theory, as in the *Trolley*-examples of the preceding paragraph. On this basis, an extended thought experiment argues for giving up a commitment or some element of the current theory.

If the process of mutual adjustments is to be carried out in a systematic way, heuristic thought experiments play a crucial role as well. They are needed for exploring the results of applying the current theory to systematic variations of scenarios. Again, the many variants of *Trolley* provide an example.

4.3. Consequences

We can now turn to some debated issues in the literature on ethical thought experiments. Let us begin with the results of core thought experiments.

Firstly, the result of a thought experiment is always a commitment that is explicitly expressed in public or in one’s mind, not one merely revealed in action – for the thought experiment must not rely on the scenario being real. It has been argued that this is problematic and we should rather rely on how we act in real situations (Davis 2012). As a general strategy to determine what is morally right, this is implausible because, regrettably, we cannot count on our actions being more reliably right than our explicit commitments. The method of reflective equilibrium, however, calls for initially recognizing and also for possibly revising all kinds of commitments, including those resulting from thought experiments and those merely revealed in action.

A second issue concerns the role of intuitions. We have seen in Section 2 that an implausibly narrow conception of thought experiments results if one insists that all core thought experiments elicit intuitions. Many further points about the role of intuitions depend on a theory of intuitions, but some basic results are available independently (for discussion, see Stich and Tobia, this volume; Ludwig, this volume; Copp 2012; Brun 2014). On the relatively uncontroversial assumption that intuitions cannot be based on explicit inference, the method of reflective equilibrium leaves room for intuitions as results of core thought experiments, as long as they are not gained by explicit reasoning; this excludes appealing to the theory we are currently developing when going through the scenario. And the conclusion of an extended thought experiment can be an intuition only if it is not reached by explicit argumentation. Furthermore, reflective equilibrium does not exempt intuitions from critical evaluation. They can be revised just as any other commitment. Sometimes, we will not accept an intuition elicited by a scenario as its proper result because we have good reason to think that the intuition is inadequate. Someone may hold that changing the switch is impermissible even if he or she is not able to get rid of his or her recalcitrant intuition that doing so is permissible in this particular case. In other cases, epistemic background theories may give us reason to discredit intuitions as based upon, say, cultural stereotypes.

Closely related considerations also make clear that requiring commitments to be considered judgements poses no principled obstacles to the use of thought experiments. It just means that the conclusion of the extended thought experiment depends on the (possibly implicit) assumption that the result of the core thought experiment meets certain standards.

Turning to extended thought experiments, we can first note that although destructive thought experiments can play an important role, this does not mean that the result of a core thought experiment always “wins” against incompatible commitments or elements of the current theory. It is rather one of the characteristics of reflective equilibrium that revisions may well take the other direction and lead to overriding the result of a thought experiment (as in the example of the recalcitrant intuition above). A refuting thought experiment can show the need for a revision, but it alone cannot settle what is to be revised.

A second point concerns analogies. It is sometimes assumed that supportive thought experiments paradigmatically take the form of analogies (see Section 5.2). Take *Violinist* as an example: is its main point not that the result of the core thought experiment can be transferred by analogy to cases of abortion? In the context of the method of reflective equilibrium, however, the standard way of transferring a moral judgement from one case to another is not by direct analogy but by explicitly introducing principles that cover both cases.

The most important consequence is more general. If we rely on the method of reflective equilibrium, then the fact that a proposition is the result of a core thought experiment is not sufficient for this proposition to be epistemically justified to the degree required for knowledge (and the same is true a fortiori for conclusions of analogies based on such a result). If reflective equilibrium is necessary for epistemic justification, the primary object of justification is not a single proposition but an entire position consisting of the commitments and the theory in equilibrium. Hence, thought experiments may play important epistemic roles according to the method of reflective equilibrium, but they cannot by themselves justify a commitment or a theory. Defenders of reflective equilibrium should therefore object to the view that a destructive thought experiment alone may suffice to refute a claim and to the idea that being supported by an intuition elicited by a thought experiment or by an analogy based on a core thought experiment may be sufficient for justification.

5. Challenges to ethical thought experiments

Although ethical thought experiments are frequently challenged, wholesale criticism of ethical thought experiments in all their functions is implausible. There are, however, challenges which target a specific function of thought experiments. Well-known examples are Dancy's attacks on supportive thought experiments, which are briefly discussed below. But most challenges to ethical thought experiments raise a problem that may affect specific thought experiments, and I focus on such challenges in what follows. I first give an overview in Section 5.1 and then discuss three issues which have been prominent in debates about ethical thought experiments (Section 5.2). Finally, I address the question of how thought experiments in ethics can raise specifically moral problems (Section 5.3).

5.1. A general overview of challenges

The structure (1)–(3) from Section 2 can be used as a basis for an overview of challenges which address specific thought experiments in any of the functions distinguished in Sections 3 and 4 (see also Gendler 2000, 22–4). Challenges to core thought experiments frequently point out a problem with the scenario or the question. Sometimes we have difficulties going through the described situation because the scenario is incoherent. In other cases, the situation is difficult to imagine because it is underdescribed; that is, described too schematically and relevant details remain undetermined. Or the scenario in effect undermines the conditions for a meaningful application of the concepts used in the question (this kind of objection is best known from the debate about personal identity, targeting, e.g., Parfit's 1987 teletransportation scenarios; see Rescher 2005, ch. 9). Two points which will be discussed further below are the question of how realistic scenarios should be, and the charge that a scenario or a question is misleading.

Other challenges to core thought experiments point out problems with "going through" the scenario and coming up with a result. Maybe no intuition or judgement results because the situation is underdescribed. In such cases, the problem is not that we cannot imagine a situation as described, but that the scenario does not provide enough information to answer the question in a non-arbitrary way (Wilkes 1988, e.g., claims that this problem arises for personal-identity thought experiments because “person” is not a natural-kind term). In other cases, an intuition or a judgement results, but in a problematic way. Whether thought experiments are unreliable if they rely on intuitions is currently a much debated issue (see Stich and Tobia, this volume; specifically for moral intuitions: Burkard 2012; Copp 2012). Many writers hold that intuitions are not reliable truth-trackers, but frequently the product of prejudice, tradition or stereotypes. One strand of experimental philosophy is devoted to investigating empirically whether such accusations can be substantiated by showing that people’s reactions to a scenario co-vary with, for example, cultural factors (see Stich and Tobia, this volume; Ludwig, this volume). If the result of a core thought experiment is arrived at not intuitively but discursively, the reasoning may be criticized as fallacious if it involves an invalid or weak argument, an unwarranted assumption (invited, e.g., by an underdescribed scenario), or a conclusion that is irrelevant in the context at hand.

Challenges to an extended thought experiment attack the reasoning which is supposed to establish the conclusion about the target. One type of criticism is that the reasoning is fallacious in one of the ways just mentioned. Since thought experiments using arguments from analogy are quite common in ethics, such objections often home in on weak analogies (see the discussion below). Other challenges point out that it is not clear enough what the conclusion of the thought experiment is supposed to be or that it misses what is at issue in a given debate. For example, does a refuting thought experiment address utilitarianism in general or only some specific version?

5.2. Common challenges to ethical thought experiments

Let us now turn to three types of challenges which are frequently raised in debates about ethical thought experiments. A first question is whether thought experiments, especially epistemic and heuristic ones, should be realistic. The issue is not whether the scenario gives a vivid and detailed description, nor whether the described situation is improbable or bizarre, but rather whether the described situation is modally remote, involving, for example, biologically or technologically impossible elements such as the fission of humans.

On the one hand, there are reasons for appealing to unrealistic situations. They are relevant to universally valid statements and definitions (Sorensen 1992, 279; Walsh 2011). They can produce results that we have a lot of confidence in (Parfit 1987, 200). And if we want to use heuristic thought experiments to isolate morally relevant factors, we need to compare similar situations which differ just in the factor in question (e.g. would cruel treatment of animals be permissible if animals were insentient?); appeal to unrealistic scenarios is then needed if rival theories agree on realistic situations or if realistic situations differ in additional respects (Sorensen 1992, 91; Kamm 1993, 7).

On the other hand, unrealistic epistemic thought experiments have been challenged on various grounds (see Elster 2011). Firstly, some approaches to ethics seek to develop practical, action guiding, moral principles which are meant to lead to the right results in our world, but not necessarily in unrealistic situations. Unrealistic thought experiments are then irrelevant to the justification of these principles (e.g. Hare 1981, 47–9). However, insofar as an approach to ethics must also deal with “deeper” ethical principles that capture moral truths or definitions of moral terms, unrealistic cases remain relevant.

Secondly, there is the worry that unrealistic scenarios lead to core thought experiments with unreliable results. This challenge needs qualification. Reliability is independent of whether the scenario is realistic overall. Unrealistic problems can have easy answers (If binary fission is iterated three times, how many persons result?), realistic cases can pose hard questions and real situations can include too many factors which distract from important features, trigger biases or interfere with our judgement in other ways. More plausibly, unreliability looms if the unrealistic aspects undermine the resources needed to answer the question at stake (Rescher 2005, ch. 9). Even if the scenario is coherent, we may not be able to determine the morally relevant factors and their exact import, or to work out the necessary background assumptions for the unrealistic situation, especially if the unrealistic aspects concern central elements of our web of belief (Sorensen 1992, 43–4). With respect to unrealistic thought experiments that ask for an intuition, many argue that our capacity for intuitive judgement is unreliable in situations for which it has not been developed by evolution or social practice (e.g. Kitcher 2012). This line of argument sometimes leads to more general scepticism about epistemic thought experiments involving intuitions (see, e.g. Singer 2005; Machery 2011).

As a second type of challenge, many epistemic thought experiments are accused of being misleading (Wood 2011), independently of whether they are realistic. At least three charges can be distinguished. Firstly, the really important moral questions are replaced by dubious problems that do not occur in reality. Instead of, say, "Who is responsible for the trolley disaster?", the experimenter asks "Should you throw the switch?" Secondly, thought experiments use forced-choice questions which artificially limit the range of admissible courses of action ("Change the switch, yes or no?") and thereby rule out giving the right answers (e.g. "Try to save all six."). Thirdly, problematic assumptions are introduced, especially by way of implicature or presupposition. *Trolley,* for example, suggests that it asks an important question, that at least one of the available answers is morally legitimate, that moral decisions are a matter of ranking states of affairs resulting from actions, and that further information about the situation is irrelevant (e.g. "Do the workers have permission to be on the track?").

However, those who advance such objections must deal with the rejoinder that they do not take seriously the thought experiments they reject, either by refusing to deal with them, by filling in details that favour their own views or by insisting on changing the subject. Moreover, unrealistic epistemic thought experiments cannot simply be dismissed as irrelevant if they target a universal moral statement or definition which covers such situations. But it is true that they are often not worked out in sufficient detail and need reconstruction. In *Trolley,* for example, the intuition or judgement that we should change the switch may contribute to refuting the view that numbers are never decisive, though it certainly does not show that consequentialism is right, that numbers always count or that rights and entitlements are unimportant. Suggesting that such conclusions follow is bad rhetoric – just as much as insinuating without further evidence that users of trolley cases draw or implicate such conclusions. Finally, a lot of the scepticism may stem from the worry that scenarios such as *Trolley* or *Ticking Bomb* are used to argue directly for a conclusion about some other situation. Typically such a move involves an analogy, and this leads to a third type of challenge to epistemic thought experiments.

Violinist and *Pond* are routinely interpreted as analogies in which the result of the core thought experiment – "It is permissible to unplug yourself from the violinist." or "You ought to rescue the child." – supports a conclusion about abortion or poverty relief respectively. In Section 4.3, I explained that in the context of method of reflective equilibrium such supportive functions are usually better reconstructed as operating more indirectly. But if extended thought experiments are interpreted as analogies, two sorts of objections need to be considered.

Firstly, Dancy (1985, 1993, 64–6) has mounted an attack on supportive thought experiments, which he assumes to employ an analogy from the result of a core thought experiment to a real case. According to Dancy, such analogies are not reliable because, in contrast to the real case, the information on the situation described in the scenario is inevitably limited. We must therefore either qualify our conclusion about the real case with the condition that it does not relevantly differ from the situation described in the scenario, or admit that any new information about the real case can spoil the analogy. One answer operates on the metaethical level and argues against Dancy's specific form of particularism and holism about reasons, which underwrites his claim that we cannot conclusively evaluate situations which are neither real nor variations of real situations (see Häggqvist 1996, ch. 2.7; Cohnitz 2006, ch. 4.2.2 for discussion). Another answer challenges Dancy's understanding of the analogy involved. He asks for an argument that, given the result of the core thought experiment, conclusively shows a claim about the real case. Arguments from analogy, however, are a form of plausible reasoning resting on a premise which states that the real case is relevantly similar to the situation described in the scenario. Analogies are therefore not deductively valid, and strong analogies can be overthrown by further arguments. Nonetheless, for supporting, in contrast to proving, a conclusion a strong analogy suffices (Rescher 2005, 59).

Objections of the second sort target weak analogies in specific extended thought experiments. Examples abound in the debate about thought experiments which use a ticking-bomb scenario to argue that torture may be morally permissible in real life. Critiques of such arguments very often point out that even if torture should be permissible in *Ticking Bomb*, real-life situations differ from this scenario in ways that effectively undermine the analogy. For example, the scenario (explicitly or implicitly) assumes that we know that only torture will get us the right information in time, but in reality this is uncertain, or that torture is strictly limited to some very specific and rare exceptions, whereas in reality it is always an institutionalized practice with a strong tendency to spread (Luban 2005).

5.3. Moral objections to ethical thought experiments

If an ethical thought experiment in fact suffers from one of the shortcomings discussed so far, this can obviously be morally relevant. Bad ethical thought experiments can lead to bad ethics and consequently to morally wrong action. Accordingly, Parfit's ethical project, for example, has been criticized for employing misleading thought experiments (Wood 2011). And attempts to legitimize post-9/11 torture practice with the help of ticking-bomb scenarios have been criticized as irresponsible because they involve an analogy which is invalidated by discrepancies between the scenario's assumptions and easily knowable empirical facts (Luban 2005).

There is, however, also the worry that some ethical thought experiments raise moral problems in other ways. Specifically, it has been argued that discussing certain scenarios and questions is a symptom of moral corruption or leads to moral corruption of the author or the audience of the thought experiment (see, e.g., Anscombe 2005; Williams 1972, 92). Such worries can hardly be defended as a general critique of (unrealistic) ethical thought experiments. After all, one may also argue that the debate about ticking-bomb scenarios, and especially the analysis of the disanalogies involved, has considerably contributed to a better understanding of how torture might not be justified. What rather seems really problematic is the rhetorical use of thought experiments such as *Ticking Bomb* in political discourse and propaganda (see Luban 2005 for an analysis and examples). Glossing over crucial assumptions and selling gerrymandered questions as well-founded challenges can have morally bad consequences and it is incompatible with an adequate self-understanding of a philosopher.

Acknowledgments

For helpful comments, I thank Christoph Baumberger, Claus Beisbart and Anna Goppel. My research has been supported by the Swiss National Science Foundation (project 150251).

Notes

[1] There is also reasoning *per impossibile* with counterfactuals with logically or mathematically false antecedents (see Rescher 2005, ch. 8.8), but as far as I can see, it plays no role in ethics.

[2] Cappelen (2012) even argues that there are no philosophical thought experiments that rely on eliciting intuitions as opposed to reasoning; he specifically discusses *Violinist* and *Trolley*.

[3] This is a rather narrow sense of "epistemic." In a wider sense, heuristic functions may also be classified as epistemic.

[4] These two ideas can be found in all standard accounts of reflective equilibrium, specifically in Rawls (1999) and Daniels (2011). In what follows, I rely on an understanding which also draws on ideas of Elgin (1996); see Brun (2014) for a more extensive outline.

References

Anscombe, Gertrude Elizabeth Margaret. 2005 [1957]. "Does Oxford Moral Philosophy Corrupt Youth?" In *Human Life, Action and Ethics.* Exeter: Imprint Academic. 161–167.

Borsboom, Denny; Gideon J. Mellenbergh; Jaap Van Heerden. 2002. "Functional Thought Experiments". *Synthese* 130, 379–387.

Brown, James Robert. 1991. *The Laboratory of the Mind. Thought Experiments in the Natural Sciences.* London: Routledge.

Brown, James Robert; Yiftach Fehige. 2014. "Thought Experiments". In *Stanford Encyclopedia of Philosophy*. (http://plato.stanford.edu/archives/fall2014/entries/thought-experiment/)

Brun, Georg. 2014. "Reflective Equilibrium without Intuitions?" *Ethical Theory and Moral Practice* 17, 237–252.

Burkard, Anne. 2012. *Intuitionen in der Ethik.* Münster: Mentis.

Cappelen, Herman. 2012. *Philosophy without Intuitions.* Oxford: Oxford University Press.

Cohnitz, Daniel. 2006. *Gedankenexperimente in der Philosophie.* Paderborn: Mentis.

Copp, David. 2012. "Experiments, Intuitions and Methodology in Moral and Political Philosophy". In Shafer-Landau, Russ (ed.). *Oxford Studies in Metaethics. Vol. 7.* Oxford: Oxford University Press. 1–36.

Dancy, Jonathan. 1985. "The Role of Imaginary Cases in Ethics". *Pacific Philosophical Quarterly* 66, 141–153.

Dancy, Jonathan. 1993. *Moral Reasons.* Oxford: Blackwell.

Daniels, Norman. 2011. "Reflective Equilibrium". In *Stanford Encyclopedia of Philosophy*. (http://plato.stanford.edu/archives/spr2011/entries/reflective-equilibrium/)

Davis, Michael. 2012. "Imaginary Cases in Ethics. A Critique". *International Journal of Applied Philosophy* 26, 1–17.

Dennett, Daniel C. 1980. "The Milk of Human Intentionality". *The Behavioral and Brain Sciences* 3, 428–30.

Elgin, Catherine Z. 1996. *Considered Judgment.* Princeton: Princeton University Press.

Elgin, Catherine Z. 2014. "Fiction as Thought Experiment". *Perspectives on Science* 22, 221–241.

Elster, Jakob. 2011. "How Outlandish Can Imaginary Cases Be?" *Journal of Applied Philosophy* 28, 241–258.

Foot, Philippa. 2002 [1967]. "The Problem of Abortion and the Doctrine of the Double Effect". In *Virtues and Vices and Other Essays in Moral Philosophy.* Oxford: Clarendon Press. 19–32.

Gendler, Tamar Szabó. 2000. *Thought Experiment. On the Powers and Limits of Imaginary Cases.* New York: Garland.

Gendler, Tamar Szabó. 2007. “Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium”. *Midwest Studies in Philosophy* 31, 68–89.

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

Häggqvist, Sören. 2009. “A Model for Thought Experiments”. *Canadian Journal of Philosophy* 39, 55–76.

Hare, Richard Mervyn. 1981. *Moral Thinking. It’s Levels, Method, and Point.* Oxford: Clarendon Press.

Hills, Alison. 2009. “Moral Testimony and Moral Epistemology”. *Ethics* 120, 94–127.

Kamm, Frances Myrna. 1993. *Morality, Mortality. Vol. 1. Death and Whom to Save from It.* New York: Oxford University Press.

Kitcher, Philip. 2012. “The Lure of the Peak”. *New Republic* 243/1, 30–35.

Luban, David. 2005. “Liberalism, Torture, and the Ticking Bomb”. *Virginia Law Review* 91, 1425–1461. (http://scholarship.law.georgetown.edu/facpub/148)

Machery, Edouard. 2011. “Thought Experiments and Philosophical Knowledge”. *Metaphilosophy* 42, 191–214.

Parfit, Derek. 1987. *Reasons and Persons.* Oxford: Clarendon Press.

Parfit, Derek. 2011. *On What Matters. Vol. 1.* Oxford: Oxford University Press.

Rawls, John. 1999. *A Theory of Justice. Revised ed.* Cambridge, MA: Belknap Press.

Rawls, John. 2001. *Justice as Fairness. A Restatement.* Cambridge, MA, Harvard University Press.

Rescher, Nicholas. 2005. *What If? Thought Experimentation in Philosophy.* New Brunswick: Transaction.

Shue, Henry. 1978. “Torture”. *Philosophy and Public Affairs* 7, 124–143.

Singer, Peter. 1972. “Famine, Affluence, and Morality”. *Philosophy and Public Affairs* 1, 229–243.

Singer, Peter. 2005. “Ethics and Intuitions”. *The Journal of Ethics* 9, 331–352.

Singer, Peter. 2009. *The Life You Can Save. Acting Now to End World Poverty.* London: Picador.

Smart, J.J.C. 1972. “An Outline of a System of Utilitarian Ethics”. In Smart, J.J.C.; Bernard Williams. *Utilitarianism For and Against.* Cambridge: Cambridge University Press. 1–74.

Sorensen, Roy A. 1992. *Thought Experiments.* Oxford: Oxford University Press.

Thomson, Judith Jarvis. 1972. “A Defense of Abortion”. *Philosophy and Public Affairs* 1, 47–66.

Thomson, Judith Jarvis. 2008. “Turning the Trolley”. *Philosophy and Public Affairs* 36, 359–374.

Uhlmann, Eric Luis; David A. Pizarro; David Tannenbaum; Peter H. Ditto. 2009. “The Motivated Use of Moral Principles”. *Judgment and Decision Making* 4, 476–91.

Walsh, Adrian. 2011. “A Moderate Defence of the Use of Thought Experiments in Applied Ethics”. *Ethical Theory and Moral Practice* 14, 467–481.

Wilkes, Kathlen V. 1988. *Real People. Personal Identity without Thought Experiments.* Oxford: Clarendon Press.

Williams, Bernard. 1972. “A Critique of Utilitarianism”. In Smart, J.J.C.; Bernard Williams. *Utilitarianism For and Against.* Cambridge: Cambridge University Press. 75–150.

Wood, Allen. 2011. “Humanity as an End in Itself”. In Parfit, Derek. *On What Matters. Vol. 2.* Oxford: Oxford University Press. 58–82.