warwick.ac.uk/lib-publications

Thought Experiments[1]

Kimberley Brownlee and Zofia Stemplowska

1. Introduction

A thought experiment is, in one sense, just what its name suggests – an experiment in thinking. But it is thinking of a distinctive, imaginative kind that offers a potentially powerful investigative and analytic tool in mathematics, science, and philosophy. In science, thought experiments are a well-accepted, uncontroversial mechanism for testing hypotheses, and in mathematics, they are one of the principal tools for valid reasoning. They can build and destroy arguments. In negative terms, they can 1) expose a contradiction, 2) undermine a key premise, 3) reveal a conflation of concepts or principles, or 4) highlight the counterintuitive implications of an argument. In positive terms, they can 1) demonstrate the consistency or coherence of a set of principles/concepts, 2) highlight congruities and similarities between different claims, 3) reveal the scope of the application of a given principle, and 4) bring forth intuitions not previously considered, amongst other things.

In philosophy, some thought experiments are highly influential, even famous. In moral and political theory, famous examples include the following:

1. Philippa Foot's/Judith Jarvis Thomson's **Trolley Problem**: A runaway trolley is hurtling toward five people who are working on the railroad track up ahead. The driver can either continue onto the track ahead, thereby killing the five, or steer onto a second track off to the side on which only one man is working, thereby killing the one. Is it permissible to turn the trolley? (Foot 2002: 23; Thomson 1985: 1395).

This thought experiment serves various purposes. Some argue, for example, that it is permissible to turn the trolley because the negative duty not to kill the five outweighs the negative duty not to kill the one. This experiment is discussed in a variety of political theory contexts, including in just war theory on the legitimacy of defensive harm.

2. Philippa Foot's/Judith Jarvis Thomson's **Trolley Problem and Transplant Case**: A doctor can save the lives of five dying patients by killing one healthy person and giving her organs to the five. If it is permissible to turn the trolley, is it also permissible to kill the one to save the five? (Foot 2002; Thomson 1985: 1396).

This thought experiment also serves various purposes. Some thinkers argue, for example, that unlike the trolley problem, it is impermissible to kill the one healthy person in the transplant case because the positive duty to save the five outweighs the negative duty not to kill the one.

3. G. A. Cohen's **Camping Trip:** Suppose that a group of friends go camping together. They could either contribute according to their abilities and resources, with the expert fisher going fishing and the skilled forager finding apples, and so on. Alternatively, they could each assert their rights over their own equipment and talents and use them to bargain with the other members of the camping group. Is it better to base a camping trip on the principles of market exchange and private ownership or on the socialist principles of collective ownership and planned mutual giving? (Cohen 2009, 3ff.).

This thought experiment is used by G.A. Cohen as his starting point to challenge the idea that socialism is infeasible and counter-intuitive, and to consider whether societies are relevantly distinct from camping trips, or whether societies that cultivate the mechanisms to harness human generosity could be governed by socialist principles.

4. John Rawls's **Original Position**: We are asked to put ourselves in the position of free and equal persons collectively deciding upon and committing to a set of principles of justice for society. To ensure our impartiality as deliberators, we engage in this process behind a 'veil of ignorance' that shields us from the knowledge about who we will be in the society and what advantages, and disadvantages we will have. What principles would we choose? (Rawls 1971).

This thought experiment is used to offer support for a liberal-egalitarian conception of *justice*.

5. Robert Nozick's **Experience Machine**: Suppose that you could put yourself into an experience machine for the rest of your life, which would give you all the experiences you find enjoyable and valuable without your knowing, once you were in the machine, that these experiences were not real. Would entering the machine be a good choice from the point of view of wellbeing? Would it give you all that mattered? (Nozick 1974: 42-5).

This thought experiment is used to challenge the claim that the realm of value is exhausted by hedonistic pleasure. Nozick argues that the realm of value and wellbeing is not exhausted by the pleasure of 'experiences'; we care about whether our experiences are real.

6. Robert Nozick's **Utility Monster**: Suppose there is a person who gets enormously greater gains in utility from any sacrifice of others than those others lose in utility through their own sacrifice. Is it morally required or morally permissible to sacrifice these persons for the sake of the monster's greater utility? (Nozick 1974: 41).

This thought experiment is used to challenge the intuitiveness of hedonistic utilitarianism.

7. **Ticking Bomb**: Suppose that you are an investigator who has a suspect in custody who you know has planted a bomb somewhere in your city. The bomb will soon explode. Suppose also that torturing this person will give you the information you need to locate and neutralize the bomb. Is it either morally permissible or required to torture the person in custody? (adapted from Walzer 1973, 166-7).

This thought experiment is used to gauge the permissibility of torture. (See the Appendix for other examples of thought experiments.)

Unlike in mathematics and science, in political theory and in philosophy in general, the use of thought experiments is a matter of lively controversy. Two especially pressing objections against their use are the following:

1. The Objection of Bias: Thought experiments both invite systematic bias and entrench existing biases;

2. The Objection of Inherent Ambiguity: Thought experiments often are inherently ambiguous, leading to inescapably opaque judgements.

These objections are troubling because they challenge the very possibility of making logically and philosophically respectable use of thought experiments. Neither objection is forceful in its general form because, if it were, it would impugn the less controversial use of thought experiments in mathematics and science and not just philosophy. These two objections, however, may be thought to target the use of thought experiments in sub-disciplines of philosophy such as political theory and moral theory where thought experiments are deployed not only for conceptual and logical purposes, but also for normative and evaluative purposes. Using thought experiments in political and moral theory in particular may seem suspect because such thought experiments abstract away from and idealize real-life cases or even invent fantastical scenarios, but nonetheless purport to guide real-life behaviour.

4

The use of thought experiments in political and moral theory is also subject to further, less weighty objections. First, such thought experiments are said to be in poor taste since they often involve fantastic scenarios of suffering, death, and cruelty that trivialize that suffering. (Ironically, thought experiments of the past were also criticized for being too trivial: see Winch 1965: 199-200.) Second, they are said to impoverish our understanding of urgent problems, as they are devoid of rich social context (O'Neill 1986: 12, 20-21 – but note that O'Neill herself does not attribute this feature to thought experiments in general, only in some contexts within political and moral theory). Third, thought experiments, such as the Ticking Bomb, are said to misrepresent the vast majority of relevant real-life cases and thus create the false impression that the world is simpler and more manageable than it is. These latter three objections can be set aside, however, because their force, while somewhat doubtful, could be granted without abandoning the practice of thought experiments in political and moral theory. They seem to invite theorists to engage in careful and tactful delineation of thought experiments rather than to abandon them altogether.

The main purpose of this chapter is to provide a guide for the use of thought experiments in political theory (although what we say holds more generally for normative theory). As part of that objective, we aim to refute the Objection of Bias and the Objection of Inherent Ambiguity against thought experiments in this area. A further, related purpose of this chapter is to flesh out positively the distinctive argumentative value that thought experiments have in normative theory. We begin by distinguishing the concept of a *thought experiment* from things with which it is sometimes conflated, namely, introspective *psychological experiments* and other argumentative tools that appeal to the workings of the imagination such as *descriptive hypothetical examples* (section 2). We then respond to the Objection of Bias and the Objection of Inherent Ambiguity, first by articulating and defending a set of necessary, formal conditions for formulating well-posed thought experiments in normative theory (section 3), and second by showing that these conditions do not preclude the use of thought experiments that involve practical impossibilities or imaginatively opaque components (section 4). We

conclude by highlighting the key 'how to' instructions for designing thought experiments in political and moral theory (section 5).

2. Definitions

Thought experiments have been characterized variously as devices 'of the imagination used to investigate the nature of things' (Brown 2007), picturesque arguments (Norton 1996: 334), purely mental procedures that aim to reveal something about the relationship between two or more variables (Sorensen 1992, 186, 205), and judgments about what would be the case if the particular state of affairs described in some imaginary scenario were actual (Gendler 1998, 398). Our conception of *thought experiments* in normative theory is as follows:

> A *thought experiment* is a multi-step process that involves (1) the mental visualization of some specific scenario for the purpose of (2) answering a further, more general, and at least partly mental-state-independent question about reality.[2]

The reference here to 'mental visualization' highlights the imaginative quality of thought experiments. They are not purely abstract or formal operations of thought. Rather, they are operations of thought structured to invite visualization. This does not mean that, by their nature, thought experiments cannot intelligibly and profitably deploy concepts that defy visualization, such as a square circle, a world with different laws of nature, or an episode of giving birth to oneself. Rather, the point in highlighting the visual quality of thought experiments is to note that they are not carried out purely at the level of abstract principle, but instead invoke particulars that are broadly irrelevant to the generality of the conclusion to be drawn from their use.

2.1 Descriptive Hypothetical Examples versus Thought Experiments

---

[2] We do not mean to settle the debate between expressivists/non-cognitivists on the one hand and cognitivists on the other. Even if normative judgments are ultimately entirely a matter of affective states (and hence are not mental-state-independent) we mean to signal that thought experiments aim to provide answers that at least appear to be partly mental-state independent.

The reference to 'mental visualization' above should not obscure the fact that thought experiments are only a subset of a broader category of hypothetical scenarios that involve visualization and imagination. A second subset of that category is *descriptive hypothetical examples*, which, unlike thought experiments, neither test nor contribute an independent step to a chain of reasoning. Purely descriptive hypothetical examples, such as 'I have in mind here someone like Anna Karenina', or 'God is an example of a perfect being', or 'Annette is a person who is so poor that her cupboard is bare', are elucidatory not argumentative. Descriptive hypotheticals and thought experiments have different functions. The former set the parameters of the type of problem under consideration and/or clarify the concepts at issue. The latter either are independent argumentative moves or test, and hence support or undermine, argumentative moves.

Although we shall not examine descriptive hypothetical examples in what follows, it is worth noting two features of them in relation to thought experiments. First, descriptive hypotheticals can be proto-thought experiments, that is, sometimes they can be easily developed into thought experiments. For instance, once we begin to describe Annette's situation to specify the type of poverty that we wish to examine, we can also use that description to test the acceptability of various responses to her plight. Hence, we might ask 'Would we be prepared to leave someone so impoverished to struggle on her own?' Our initial description of Annette's impoverishment is not a thought experiment, but it opens up the prospect of posing questions about how to treat Annette.

Second, unlike thought experiments, descriptive hypotheticals can assume what they are meant to illustrate. This would be a fatal problem for a thought experiment as part of a chain of reasoning, but not for a descriptive hypothetical as an elucidatory device. We return to this in section 3.1.1 below.

2.2 Psychological Experiments versus Thought Experiments
The second part of our conception of a *thought experiment* - that its function is to answer a further, more general, and at least partly mental-state-independent question about reality - allows us to distinguish thought experiments from *introspective psychological experiments* (or 'psychological experiments' for short). The latter are mental procedures

that aim simply to predict or to reveal to us our psychological/mental states. A psychological experiment asks such things as: 'Can you make yourself believe that you are an oyster?'; 'Can we imagine what it is like to be a bat?'; 'Putting aside whether it is permissible, would we actually be able to bring ourselves to turn the trolley?'; 'How would you feel if your child were killed?' Psychological experiments are a distinctive kind of mental experiment in which the generation of a given mental state is precisely and uniquely what is being tested. For instance, when you ask someone whether, in circumstances C, she would fear an attacker enough to kill him, your aim is to ascertain through this test what her mental state is likely to be in such circumstances (or at least what she thinks it is likely to be). By contrast, when you ask an accountant what is 1236 divided by 3, ascertaining her mental state is not normally the object of the 'experiment' (unless you wish to find out how she will react). The object of the experiment is to get at some feature of the world – the answer 412 – that is independent of her mental state.

The commonly asserted claim that thought experiments, such as Ticking Bomb, generate strong intuitions invites confusion between thought experiments and psychological experiments because it can be read to imply that all that thought experiments are meant to test are affective (psychological) states. But the confusion between thought experiments and psychological experiments may also have a deeper source in that some thought experiments necessarily include psychological experiments as a preliminary step in order to reach further conclusions. This occurs when (and because) the variables that a given thought experiment examines include or depend upon psychological states, usually ones involving emotions (i.e. affective states). For example, take the following thought experiment:

> **Attacker:** Suppose that we see a person being attacked. And suppose that we are morally required to call the police when we see a person being attacked. If the police cannot arrive in time, are we also morally required ourselves to kill the attacker (assuming that our action will not threaten the institution of policing)?

In order to engage with this thought experiment, it may be necessary amongst other things to run a psychological experiment by asking ourselves if we would be able to bring ourselves to kill the attacker. (Would you?) We might want to ask this question if, say, we accept that 'ought implies psychological can', i.e. if we were psychologically unable to bring ourselves to kill the attacker, then, if ought implies psychological can, we would not be morally required to do so. Nonetheless, although this psychological experiment is part of 'Attacker', that thought experiment is not exhausted by the performance of the psychological experiment since the thought experiment requires us in addition to reach a *judgement* about a moral requirement. That is, it requires us to reach a judgement about what is morally required of us in this kind of case (and that judgement is, on standard objective conceptions of morality, at least partly independent of our beliefs as the agent).

Given that psychological experiments in political and moral theory usually test affective states, one rough and ready way to distinguish thought experiments from psychological experiments in this area is to think of thought experiments as answering 'What is your moral judgement?' and psychological experiments as answering 'How would you feel?'[3] We stress the distinction between thought experiments and psychological experiments for several reasons. First, it allows us to explain the nature of the intuitions that thought experiments are designed to elicit. Simply put, unlike most psychological experiments, thought experiments are not intended to elicit raw, unreflective intuitions or brute reactions. Their results can and often should be the fruit of reflection. Second, it has implications for the way in which data gathered through thought experiments can enter into normative reasoning. If what matters in thought experiments are not (or not exclusively) raw affective states, then there is more room for rational debate over the appropriate response to a given thought experiment.

2.3 Simple Thought Experiments versus Complex Thought Experiments

---

[3]   One might worry that an expressivist or non-cognitivist would reject this distinction is a false one. But a sufficiently sophisticated version of non-cognitivism presumably accepts that, even if moral judgment is ultimately a matter of affective states, there is nonetheless a plausible distinction to be drawn between raw affective states and 'gardened' or reflective ones.

Within the category of *thought experiments*, there are some important distinctions to be drawn. The first is between simple and complex thought experiments. A simple thought experiment, such as the Trolley Problem, considers a single scenario. In political and moral theory, simple thought experiments tend to raise questions of whether some action is morally wrong, permissible, or obligatory, or whether some state of affairs is fair, equal, just, good, and so on. For instance, the Trolley Problem raises the question of whether it is permissible to turn the trolley and divert the harm from the five to the one. Oftentimes, the theorist's intuitive, though not unreflective, response to such thought experiments is taken to be *evidence* for or against the hypothesis being tested in the thought experiment (e.g. that turning the trolley is morally permissible/required).

By contrast, a complex thought experiment, such as the combination of the Trolley Problem and Transplant Case, considers two or more scenarios in relation to each other.[4] This complex experiment – Trolley Problem and Transplant Case – contrasts (the simple thought experiment) Trolley Problem with (the simple thought experiment) Transplant Case. It aims to establish whether our normative answers in the one case align with our answers in the other case. In political theory, complex thought experiments serve various negative and positive argumentative purposes, such as exposing a disanalogy or confirming an analogy, undermining or affirming a hypothesis, revealing a conflation of concepts or principles, and bringing to light unacknowledged intuitions.

This distinction between simple and complex thought experiments is significant because some simple thought experiments need not satisfy the condition of validity (see below) that applies to all complex thought experiments.

2.4. Contingency, Necessity, and Imaginability
The final distinctions to highlight within the category of *thought experiments* relate to their degree of practical possibility and of imaginative clarity.

Thought experiments can be more or less practically possible. The category of *hypothetical* is a continuum that includes both the likely and probable though non-actual at one end, and the extremely unlikely and even the impossible at the other end. The

---

[4]   Paradigmatically, complex thought experiments involve pairwise comparisons.

former can be described as *contingently hypothetical* (e.g. 'Imagine that you are walking by a pond and spot a drowning child.'). Such hypotheticals can be structured around actual events, such as rescue boat operators in New Orleans after Hurricane Katrina having to choose between rescuing the five adults on one rooftop or the three children on another, but they remain hypothetical in that they abstract away from the real case. One reason to construct contingently hypothetical thought experiments even when real world cases are readily available is that often the mention of a real world example invites protracted discussion of the facts, which can distract from the problem at hand.

At the other end of the continuum lie thought experiments that can be described as *necessarily hypothetical* (e.g. 'Imagine a spear flying toward the edge of the universe') or at least *necessarily hypothetical for us here and now* (e.g. 'Imagine a society that has eliminated poverty'). Thought experiments that fall closer to this latter end of the spectrum are controversial to some because they depart significantly from our lived, everyday reality. Being necessarily hypothetical in either of the two senses just noted is one way in which a normative-theory thought experiment may be said to be 'wacky'.

Another way in which a normative-theory thought experiment may be 'wacky' is in being *imaginatively opaque*. Robert Nozick's Utility Monster described above involves an imaginatively opaque being since his, her or its pleasure in sacrificing others must be of a fantastic quality to outweigh the acute suffering of all of those who are sacrificed. As ordinary creatures with ordinary abilities for happiness, we are unable to imagine the kind of fantastic happiness that such a being would have to feel in order to outweigh the suffering it caused to those it sacrificed. (Parfit 1986, 389).

These two senses (and sources) of wackiness – necessary hypoetheticality and imaginative opacity – can overlap but are conceptually distinct, since some cases of practical impossibility, such as my jumping 100 feet in the air, are nonetheless readily imaginable, while some cases of imaginative opacity, such as the experiences of sleepwalking and of insanity, are nonetheless readily practically possible and indeed actual. Commonly cited examples of wacky thought experiments include Rawls's Original Position, Nozick's Experience Machine, and Nozick's Utility Monster. Since imaginatively opaque and necessarily hypothetical thought experiments invite the most

11

controversy in normative theory, they will be the main focus of our defence of thought experiments in section 4.

3. Necessary Conditions for Well-Posed Thought Experiments

What would a well-posed thought experiment in political theory look like? In this section, we outline and defend two necessary conditions for well-posed thought experiments in normative theory: philosophical respectability (section 3.1), and argumentative relevance (section 3.2). In broad terms, these conditions of well-posed thought experiments apply outside of normative theory, but they are particularly salient to normative theory. Although both conditions apply to both simple and complex thought experiments, the first condition places different constraints upon each type of thought experiment.

These conditions are a non-exhaustive set in the sense that there are further conditions that any good argument must meet (e.g. clarity), which we do not mention, as we wish to focus upon what is special to thought experiments in particular. We believe that thought experiments that satisfy these conditions will be genuinely well-posed provided they do not fall foul of conditions that apply more generally to philosophical investigation.

3.1. Philosophical Respectability

This condition has two distinct dimensions, the first of which is non-question-beggingness (section 3.1.1), which applies straightforwardly to both simple and complex thought experiments. The second dimension, validity, applies to all complex thought experiments and to some simple ones (section 3.1.2).

3.1.1. Non-Question-Beggingness

Thought experiments should not assume an answer to the question that they pose. So, when formulating thought experiments, one cannot assume that it is permissible to torture the bomb-planter in order to argue that it is permissible to torture him.

Of course, it is not always obvious when answers are being assumed and questions are being begged. Consider Wittgenstein on the ontology of moral judgments:

Suppose one of you were an omniscient person and therefore knew all the movements of all the bodies in the world dead or alive and that he also knew all the states of mind of all human beings that ever lived, and suppose this man wrote all he knew in a big book, then this book would contain the whole description of the world; and what I want to say is, that this book would contain nothing that we would call an *ethical* judgment or anything that would logically imply such a judgment (Wittgenstein 1965: 6).

Wittgenstein's scenario presupposes that there are no facts in the world of the kind that he aims to deny and hence his hypothetical does not test the claim that this is the case. A conclusion other than that reached by Wittgenstein about the content of the book would involve contradiction. This would be a fatal problem for a thought experiment testing whether ethical judgements are facts. However, the example would be acceptable as a descriptive hypothetical that merely aims to elucidate what Wittgenstein means by *facts*.

3.1.2. Validity

When we first pose a thought experiment to ourselves, we should pose it as an open question, in the way that all of the thought experiments presented above have been posed. However, the question hopefully gives rise to answers, that is, the results of the thought experiment. Results are broadly of two types. First, they may simply consist in answers about what is morally required, permissible, etc. (e.g. it is impermissible to kill one to save five in 'Transplant'). Second, they may consist in such answers together with a further hypothesis about why this is the correct answer (e.g. because harming is worse than not aiding). Simple thought experiments allow but do not require the researcher to propose the hypothesis. All complex thought experiments, however, necessarily contain at least an implicit hypothesis about what does the work in one of the simple thought experiments; the next simple thought experiment is then added precisely in order to test that hypothesis (see below).

Where a thought experiment contains or generates a hypothesis explaining our intuitive reactions to the scenarios involved, the thought experiment can be 'translated' into an argument. That argument must satisfy the condition of validity. In other words, the argument should not involve logically fallacious reasoning. Of course, what constitutes fallacious reasoning is a matter of some debate. The point is simply that thought experiments that contain and generate hypotheses can and should be held to the same standards of valid reasoning as conventional arguments, whatever those standards may be. Inability to translate such thought experiments into a valid argument would indicate that we are unsure either of what the experiment is supposed to test or of whether it presupposes what it is meant to reveal.

An example of a thought experiment that satisfies the validity condition is the following from Peter Singer.

> **The Pond and the Envelope**: Imagine that you are walking by a shallow and isolated pond in which you spot a drowning child. You can easily save the child. Must you save the child? Imagine next that you receive a letter from a charity such as Oxfam asking you for a donation that you can easily make, to save a child (or, likely, many children) abroad. If you accepted that you must save the child in the pond case, must you also save the child(ren) in the envelope case? (Singer 1972, 231-2).

The thought experiment is used to show, amongst other things, that it makes no moral difference whether the person we can help is a neighbour's child ten yards away or a foreigner whose name we shall never know, ten thousand miles away.

Assume that, following Singer, you answer all questions in the affirmative. Once we form an intuitive affirmative response to the questions, this thought experiment can be readily translated into a valid, conventional argument as follows:

P1: We can easily save the child in the pond.
P2: We have a duty to save the child in the pond.

P3: The best explanation for P2 is that we have a duty to save others when we can do so at little cost to ourselves.

C1 (the hypothesis): We have a duty to save others when we can do so at little cost to ourselves.

P4: We can send money to Oxfam at little cost to ourselves.

P5: We can save others by sending money to Oxfam.

C2: We have a duty to send money to Oxfam.

The condition of validity is satisfied here since the argument into which the complex thought experiment is translated tracks what the thought experiment was intending to test or establish and satisfies the criteria for a valid argument. More generally, of course, Singer would want us to see the argument not only as valid but also as sound: he would want us to accept that there is no relevant difference between the Pond and the Envelope scenarios in that both require the same moral response and both are explained by the same general principle (the hypothesis: C1).

A thought experiment that does not satisfy the condition of validity is the following:

A Sibling and a Stranger: Imagine that your sibling contracts malaria and can be saved only if you agree to finance expensive medical treatment involving a helicopter ride. You can finance it, albeit it will cost you a lot and you won't be able to go on holiday for a few years. Must you do it? Imagine next that your sibling is healthy but you can finance similar life saving medical treatment for a stranger. If you accepted that you must save your sibling, must you not also save the stranger?

This thought experiment can be readily translated into a fallacious argument:

P1: We can save our sibling at high cost to ourselves.

P2: We have a duty to save our sibling.

P3: The best explanation for why we have a duty to save our sibling is because this is our sibling.

C1 (the hypothesis): We have a duty to save our siblings even when this involves a high cost to ourselves.

P4: Saving strangers would involve high costs.

P5: The costs of saving the strangers would be identical to those of saving our sibling.

C2: We have a duty to save strangers even when this involves a high cost to ourselves.

This argument is invalid because, even were the premises all true, the conclusion need not be true: it does not follow from the fact that we have a duty to save a sibling at high cost to ourselves that we necessarily have a duty to do other things that are equally costly.

The condition of philosophical respectability has the virtue of demystifying the status of thought experiments. If thought experiments can be represented as conventional arguments that meet the standards of valid reasoning, then it is unsurprising that they can act as solutions to philosophical problems. As we argued, this is the case with all complex thought experiments and with at least some simple thought experiments. To be genuinely well-posed, however, thought experiments should also be analytically useful to us and not corrupt our reflections. This is addressed by the second condition, the condition of argumentative relevance.

3.2. Argumentative Relevance

Thought experiments should be designed in such a way that we can focus on the relevant aspects of the scenario. We do not want our intuitive answers to respond to features of the scenario that are not part of the test and that thereby pollute it. For example, when testing a given hypothesis (such as a hypothesis about how we ought to treat strangers), it is necessary not to construct scenarios that more plausibly test an alternative hypothesis (such as a hypothesis about how we ought to treat our siblings), as we cannot assume that

they will elicit the same answers. For instance, the Sibling and the Stranger could be translated into a valid argument that fails to meet this condition:

P1: We can save our sibling at high cost to ourselves.

P2': The best explanation for why we have a duty to save our sibling is because we have a duty to save others even at a high cost to ourselves.

C1' (the hypothesis): We have a duty to save others even when this involves high cost to ourselves.

P3: Saving strangers would involve high costs

P4: The costs of saving the strangers would be identical to those of saving the sibling

C2: We have a duty to save strangers even at a high cost to ourselves

The argument is valid, but ridiculous: P2' misidentifies the principle to be derived from considering the case of the sibling. Similarly, looking at a simple thought experiment, if, in the Transplant Case, we forbid the doctor to kill the one person to save the five on the grounds that a doctor may never kill, then we are not testing what the experiment is meant to test which, among other things, is whether there is a normatively significant difference between killing and letting die (see likewise Rivera-Lopez 2005). By prohibiting the doctor from killing, we block the relevant test, as we allow her status as a doctor to infect our reflection upon the scenario.

All in all, if we want to use our thought experiments as *evidence* for or against a given hypothesis (premise), we need to make sure that the results of the experiment actually support or challenge the argumentative move in question. The key question is whether it is possible to make this condition of testing the relevant hypothesis more concrete beyond prohibiting obvious shifts in focus. We argue that the condition of argumentative relevance translates into two weak constraints upon the design of thought experiments. The first requires that the experiment allow for rudimentary alternatives (the Rudimentary Alternatives Constraint). The second requires that the experiment not

encourage narrative-framing bias (the Moderate Narrative-Framing Constraint). These two weak constraints can be contrasted with more demanding variants, which we reject.

3.2.1. Rudimentary Alternatives Constraint

When we assume the absence of some (believed) necessary feature of the world, we should stipulate at least a rudimentary alternative. The aim here is to eliminate the bias that may come from continuing to assume that the feature still obtains. For example, an ancient philosopher who believes that objects can only move by *willing* to move, should not run a thought experiment like the following:

> **Unwilling Rock:** Assume that a large rock is not willing to move, but still moves. Is the rock to blame when it kills someone?

The ancient philosopher who holds that willing is a necessary condition for moving should not run this thought experiment - without some extra stipulations - because he is pre-committed to the view that the object that moved must have been willing to move. He should stipulate instead a rudimentary alternative for *how* the object moved; for example, the object moved because it fell just like a human being might fall if pushed by a gust of wind.

> **Unwilling Rock 2:** Assume that a large rock is not willing to move, but still moves, pushed by a gust of wind (against its will). Is the rock to blame when it kills someone?

Although we endorse the Rudimentary Alternatives Constraint, we reject the more demanding Fleshed-Out Alternatives Constraint, which holds that allowed alternatives must be fully fleshed-out and rendered comprehensible to us given what we know about the world (Wilkes 1988, 43ff.). Returning to the ancient philosopher, unlike the Rudimentary Alternatives Constraint, the Fleshed-Out Alternatives Constraint would require him to explain *how* a large, heavy rock can be pushed by a gust of wind. We

18

acknowledge that a fully fleshed-out alternative would protect us from certain biases, but the protection is too restrictive. It is implausible to hold that we really need a fleshed-out, clear statement of the alternative to the ruled-out feature of the scenario, in order to prevent the ruled-out feature from determining our conclusions.

All in all, then, we accept that unconscious bias is real bias. But the possibility of bias is not a reason to abandon theorizing that might be subject to it. It is a reason to guard against it within the parameters of the case. We think that the Rudimentary Alternatives Constraint allows us to do so.

3.2.2. Moderate Narrative-Framing Constraint

We also support, more tentatively, the Moderate Narrative-Framing Constraint that guards against thought experiments that encourage narrative-framing bias. For instance, a thought experiment that draws its scenario from a well-known novel, event, film, genre, cultural myth, or icon can bring with it considerable narrative baggage in that the context of its creation has its own purposes that might subordinate or undermine clear reflection upon the scenario as a thought experiment. The problem is best explained with the use of an example that we owe to Roy Sorensen (1992: 264; see also Parfit 1986: 199-200; Coleman 2000: 58-60). Consider teleportation. Since it is almost exclusively encountered in the context of Sci-Fi adventures such as Star Trek and Harry Potter, its context makes demands of narrative unity upon our reading of teleportation scenarios. As viewers, we want to believe, for the sake of the story, that it is the same person who is teleported rather than a new person who is created by the process, and this may infect the philosophical use we seek to make of teleportation scenarios.

Sorensen goes on to suggest less plausibly that Nozick's Experience Machine also may be systematically distorted for a similar reason, namely, that we approach this thought experiment as a story about someone entering an Experience Machine and we find the possibility of such a story so unbearably boring that we reject it as a legitimate prospect. But, Sorensen's position on this is implausible. There is no putative demand of narrative unity about Experience Machines that necessitates that an Experience Machine scenario be irretrievably boring. We can re-write this kind of scenario as an exciting,

Matrix-style adventure that eliminates the supposed anti-boredom bias. Our re-writing may introduce a pro-excitement bias in favour of the adventure, but that suggests that we need only find a middle-of-the-road description of the Experience Machine experience.

The same is true presumably for most thought experiments. Narrative bias need not hopelessly infect thought experiment scenarios provided that we are attentive to the structure of the scenario and to the narrative assumptions that it can imply. Thus, for example, when we involve the Nazis to make a point against the permissibility of medical experimentation, we should be careful not to appeal just to the horrors that the mention of Nazis invokes.

By endorsing the Moderate Narrative-Framing Constraint in the form given above, we reject the more demanding Extreme Narrative-Framing Constraint that requires thought experiments to be 'maximally conservative' and lie exclusively within the realm of *contingent hypotheticals* and never that of *necessary hypotheticals*, the spectrum of which we discussed in section 1 (see Haggqvist 1996: 147; Rivera-Lopez 2005). The central idea behind maximal conservatism is that experiments that require us to depart from standard circumstances that we would encounter in our world will not track our reactions to the features of the case *as set out in the experiment* but will instead track our reactions to the standard features of a case *encountered in the actual world*. For example, when asked to assess whether to kill one to save five in the Transplant Case, we will ultimately not be able to take on board the stipulation that the alternative deaths really are certitudes, since in our common experience we may hope that the five would still have a chance of surviving since we never know for certain. This alleged limitation of our mental abilities calls into doubt the usefulness, or even the possibility, of wacky thought experiments such as necessary hypotheticals.

But, while the problem of polluted intuitions is genuine, the assumption underlying the postulate of maximal conservatism is mistaken. The postulate rests upon the wrong-footed assumption that we are likely to reach better judgments when we operate within a familiar, real-life context than when we operate within an unfamiliar context. But this is not generally true. We might be better able to react only to the variables that the thought experiment is meant to test when the case is set in the context

that we do not normally encounter and are not familiar with, just as a non-native English speaker might be quicker than a native speaker to spot certain linguistic patterns in English, or a person unfamiliar with a given family's dynamics might be quicker than family members are to spot instances of mental abuse and exploitation. A radically unfamiliar context may well make us more attentive to the features of the scenario that matter precisely because we are less likely to smuggle in additional assumptions.

That said, we do not deny the potential legitimacy of the above worry about narrative bias. A thought experiment that asks us to assume a hateful baby or a saintly Mafioso might be hard to execute correctly. Similarly, we might also worry in relation to the ticking-bomb thought experiment that it asks us to assume what is very hard to imagine, namely that the torturer will be exceptionally well-informed and never tempted to abuse his power. However, thought experiments are processes that we can approach slowly and reflectively, thereby guarding against possible biases. If such biases occur, this does not rule out the use of thought experiments, but rather requires us to redesign them, especially as similar, if not greater, biases are likely to plague actual, real-world scenarios.

4. Why Wacky Thought Experiments Can be Well-Posed

We reject the possibility that a bar on wackiness is one condition of a well-posed thought experiment. In this section, we explain why.

The 'wackiness' of thought experiments can be disambiguated into the two main categories noted in sections 1 and 2: *imaginative opacity*, and *necessary hypotheticality*, including *necessary hypotheticality for us here and now*. We hold that neither of these dimensions of wackiness bars thought experiments from being well-posed. Before we consider imaginative opacity or necessary hypotheticality, we need to address an objection that applies not only to these dimensions of wackiness, but also to something that is not at all wacky: contingent hypotheticality.

4.1. Contingently Hypothetical Thought Experiments

Some critics seem to hold that we can conduct respectable normative theory without thought experiments. Such critics maintain that it does not matter that a given theory or principle would lead to counterintuitive recommendations when checked against a hypothetical scenario because all we need to know is whether our principles will serve us in our *actual* world (e.g. Hare 1981, 135 and chapter 8 passim; for discussion see Carson 1993). The suggestion is that all that matters in normative theory is whether our principles perform well in the actual world, and not how well they cope with non-actual situations.

Against this view, we offer two arguments. First, the performance of our principles in actual scenarios is not all that matters in normative theory. Frequently, normative principles are put forward as explanations of *why* a given course of action is right or wrong. Call such principles *explanatory principles*. Explanatory principles must be more abstract than the principles that bear upon a given actual situation if they are to explain our judgements in that situation. For instance, it is wrong to kill five patients in order to harvest their organs to save one patient because it is wrong to kill in order to save. Since such explanatory principles are more general than the situations that call for them in the actual world, they must be testable against all situations to which they could, in principle, apply. This means that they also stand or fall by their performance in hypothetical situations. This is one reason that political theorists often need to consider counterexamples, and rely upon hypothetical scenarios to expand their range, in order to challenge abstract principles.

Second, even if we are not seeking explanatory principles, but merely principles that will serve us well with the problem at hand, thought experiments offer us a way of trying to resolve disagreement on which principles these are. All else being equal, it should count in favour of a principle that it holds across a wider range of scenarios (just as it counts in favour of a theory that it applies to a wider range of problems). So, thought experiments are important in normative theory because they can help us break ties.

These responses might be taken to suggest that normative theories (call them applied theories) would not need thought experiments if they did not seek explanatory principles or were not challenged by other theories. But even this suggestion should be

resisted. That is, even applied theories that carry support need a method to discern whether the considerations brought to bear on a given problem (to reach an all-things-considered judgement) are relevant from the moral point of view. For example, we need to ask whether it is relevant to consider such things as 'this is my house', or 'this is extremely demanding', or 'she is a woman', or 'it's repellent', or 'I would not do it myself'. Thought experiments offer an efficient way of testing the salience of such considerations.

4.2. Imaginative Opacity

Turning to imaginative opacity, this dimension of wackiness raises the following worry (Parfit 1986, 389; Wilkes 1988, 15ff; Cooper 2005, 328-347). Imaginatively opaque thought experiments fail to have an adequate imaginative grip and hence they pose 'what if?' questions that the experimenter cannot answer. The reason that the experimenter cannot answer those questions may be that she has no knowledge of the laws that govern the behaviour of the entity she is imagining. Or she may have knowledge of the laws relevant for predicting that behaviour in the actual world (e.g. the process of human birth), but those laws do not apply in the hypothetical scenario (e.g. giving birth to oneself). The fact that the experimenter cannot answer these questions is said to negate whatever argumentative value the thought experiment might have.

At least two replies can be made to this objection. First, ruling out the use of imaginatively opaque thought experiments would be unduly prohibitive. It would rule out the use of thought experiments that expose certain paradoxes. For instance, it would rule out the use of a thought experiment used to show that causal paradoxes would emerge if one could go back in time and kill one's father. But the opacity of the experiment does not stop us from pointing out the potential paradoxes.

Second, it is not clear that being able to imagine all aspects of a given case is essential to run the thought experiment. For example, we (the authors of this chapter) cannot fully conceive of a being that is both a dog *and* able to talk, but we can still ask whether such a dog would count as a person. Likewise, we cannot imagine a utility monster that derives almost boundless pleasure from the suffering of others, but we can

ask whether such a being would be right to make others suffer. Recall that thought experiments are not 'run' (simply) to establish how we (the experimenters) would *feel*, but to establish what we may plausibly *think*, and hence we may not require a full character brief in the way that actors would if they were required to play a given part and to react 'in-character'.

4.3. Necessary Hypotheticals

The worry that we do not understand the laws that govern some imaginatively opaque cases resurfaces in a form that applies also to necessarily hypothetical thought experiments. The worry takes the form of a dilemma:

> Thought experiments are useless because we either cannot set them up properly or cannot derive any credible conclusions from them. That is, either we are assuming a world similar to ours, in which case we cannot set up a wacky thought experiment at all (e.g. in a world similar to ours, dogs are not as intelligent as persons; there is no teleporting; etc.), or we are assuming a world that is radically different from ours, in which case we cannot know what to say about *this world*.

(See Raz 1986: 419-20; Mulhall 2002: 16-18.)

Why can we apparently not know what to say about this world? The answer relates to 'semantic holism' (see Sorensen 1992: 282-4). The idea is that *our* concepts developed to track *our* world, rather than the wacky worlds that we set up in our experiments, and wacky worlds cannot plausibly capture our real-life concepts. For instance, suppose that half of the standard tests that we use to determine whether a piece of copper wire is electrified give us a positive answer and half of the tests give us a negative answer. What should we say about this piece of copper wire? The answer, according to Nowell-Smith should be: 'I simply do not know what I should say' (1954: 240; quoted in Sorensen 1992: 283). Now consider a wacky, normative-theory thought experiment.

Rich and Superrich: Imagine a world in which there are only rich and superrich. Is the inequality that holds between the rich and the superrich unfair or otherwise problematic?

The answer, according to a critic of wacky thought experiments, is 'I simply do not know' since our concepts developed to deal with entirely different cases and they are of no use in radically re-imagined worlds. To give an analogy, paint colours developed to paint the British landscape are of little use in painting the African landscape, given the very different light of the two environments.

However, this objection rests upon a mistake. It assumes that thought experiments ask us what we *would* say if our concepts were developed to accommodate the wacky cases as standard. But this is not what thought experiments ask us to do. They ask us, instead, to judge how our current, familiar concepts behave when exposed to new situations. To see this, first, return to the paint analogy. The thought here is that we are not asked to use the British paints to paint the African landscape; we are asked instead to use the African light to rule, say, on whether two identical-looking British colours really are identical. When we cannot easily tell if the colours are the same against the British light, we may benefit from examining them under the African light. We examine the value of equality by asking if we still value equality in such a context. If not, then we have reason to suspect that what matters to us in our ordinary context is not simply equality, but absolute levels of deprivation.

Second, consider another illustration of the application of our ordinary concepts to new situations. In Rich and Superrich, we do not ask what we would think if we were living in such a privileged society. Instead, we ask whether we consider the inequality present in that privileged society to be too unfair by our own, current standards. Pointing out that in a world with only the rich and the superrich, no one would care about equality (and that they may not even have a sense that they are unequal) is irrelevant to the question of whether we now see the inequality as problematic. (There is a wrinkle here. We may have a conception of *unfair inequality* according to which inequality is only unfair if the people subject to it consider it to be unfair; if this is so then, indeed, we may

be unable to tell whether unfair inequality characterizes the hypothetical scenario but that is not because the thought experiment is hypothetical and wacky, but because we do not have the relevant empirical data about the people we are investigating.)

To conclude this section, we want to emphasize that wacky thought experiments are not, in fact, used solely to advance academic debates, which some might consider esoteric in any case. Testing how our familiar concepts behave when exposed to new situations is a common, undisputed, and powerful strategy in many fields, including one closely aligned with normative theory, namely, Law. It is routine practice in Law Schools to hold Moot Courts revolving around wild and wonderful cases so as to train Law students in the application of key legal concepts. And the application of such legal concepts as *theft* and *property* is viewed as no less legitimate when the parties are Martians stealing magic from Venusians. Moreover, in broad terms, real court cases are exercises in thought experimenting since both ordinary court procedures and norms of due process necessarily yield an abstracted and idealized presentation of the facts of the case.

Ultimately, wacky thought experiments are not undermined by our inability either to imagine all of their elements or to anticipate how the concepts we are exploring would evolve in hypothesized worlds.

5. Conclusion: How to Design Good Thought Experiments in Political Theory

This discussion has identified several conditions for good thought experiments. We conclude here by presenting these conditions as a set of 'how to' instructions for the design of good thought experiments in normative theory:

1. We may stumble across thought experiments in film and novels; and life itself can generate situations that can be thought through as thought experiments. But, usually the impetus for designing them is that we want to test some premise in our argument/some hypothesis about a plausible principle or value. If so, **the most fundamental question you should pose is this: What exactly would you like your thought experiment to test (to undermine or to support)?**

2. **Ensure that the thought experiment is relevant to what is being tested** (see section 3.2). **Try to design the simplest experiment** in the sense of having the most parsimonious story (to avoid introducing distorting elements), but do not worry if the story is also wacky in the sense of being fantastical. (Our defence of wacky thought experiments is meant to set your imagination free).

3. **If your thought experiment involves a denial of a standard feature of the world (e.g. you deny that the police are uncertain whether they caught the right guy), hypothesize, even roughly, how this can be** – what alternative feature of the world is present. (See section 3.2.1.)

4. **Be sensitive to possible narrative-framing biases**, that is, to the structure of the scenario and to the narrative assumptions that it can imply. (See section 3.2.2.)

5. **Consider whether any imaginatively opaque elements of the thought experiment need to be imaginatively clear** in order for the thought experiment to function as intended. (See section 4.2.)

6. When constructing a necessarily hypothetical, **consider whether the selected 'foreign context' best serves to illuminate the features of our ordinary concepts and/or principles that are under consideration**. (See section 4.3.)

7. **Ensure that none of the features of the experiment already assumes what is to be tested.** (See section 3.1.1.)

8. **Ensure that the thought experiment, together with the hypothesis it is meant to support or deny, translates into a valid argument.** (See section 3.1.2.)

Appendix: Examples of Thought Experiments

1. Judith Jarvis Thomson's **Famous Violinist**: a healthy person awakes in a hospital to find that, unbeknownst to her, she has been connected by her kidneys to a famous violinist. The famous violinist will die unless she remains connected to him for the next nine months, just as a foetus will die unless it remains 'connected' for nine

months to the pregnant woman carrying it. Is the healthy person morally obligated to remain in the hospital with the violinist for nine months? (Thomson 1971: 48-9).

The thought experiment is used to challenge the impermissibility of abortion.

2.   Judith Jarvis Thomson's **Loop Trolley**: the two tracks in the Trolley Problem split, but then circle back to form a loop. Diverting the trolley onto the track with the one person will cause the trolley to hit the one and thereby prevent the trolley from continuing around the loop to hit the five. Is it permissible to turn the trolley away from the track with the five and onto the track with the one? (Thomson 1985: 1402-3).

This thought experiment is used to show that, like the Transplant Case, turning the trolley in this case uses the one person merely as a means.

3.   Judith Jarvis Thomson's **Fat Man on the Bridge:** the runaway trolley of the Trolley Problem cannot be stopped by diverting it, but can be stopped by pushing a fat man off a bridge onto the track in front of the trolley, resulting in his death. Is it permissible to push the fat man? (Thomson 1985: 1409).

This thought experiment is used to show that pushing the fat man uses him merely as a means and infringes his right not to be so treated.

4.   Bernard Williams's **Jim and the Indians**: A jungle explorer Jim comes across a colonialist Pedro who has twenty aboriginal people lined up before a firing squad. Pedro offers Jim the privilege of killing one and letting the rest go free. Jim can either refuse the offer in which case all twenty people will be killed or he can accept the offer and kill only one of them himself. What should Jim do and how should he feel about his choice? (Williams 1973: 98-100).

This thought experiment is used to question the thought that a plausible moral theory must be consequentialist. Focusing only upon the lives lost, it seems that Jim must kill the one; but Williams argues that morality is not only about good and bad states of affairs.

5. Robert Nozick's **Wilt Chamberlain** case: Suppose that basketball player Wilt Chamberlain attracts huge crowds to his games. He signs a contract with a team whereby he receives 25 cents for every home game ticket sold. During one season, a million people come to his games and happily drop 25 cents into a special box for him, thinking it well worth it. He earns $250,000 this way, a much larger sum than the average income. Is he entitled to keep all of it?

This thought experiment is used to show that people's exercises of personal freedom, such as choosing to pay 25 cents to go Chamberlain's games, will disrupt a patterned principle of distributive justice, such as the principle that everyone should have equal resources.

References:

Brown, James Robert. 2007. 'Thought experiments', in Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition). http://plato.stanford.edu/archives/fall2009/entries/thought-experiment/.

Cohen, G. A. 2009. *Why Not Socialism?* Princeton: Princeton University Press.

Coleman, Stephen. 2000. 'Thought experiments and personal identity', *Philosophical Studies* 98: 53–69.

Carson, Tom. 1993. 'Hare on utilitarianism and intuitive morality', *Erkenntnis* 39: 305-31.

Cooper, Rachel. 2005. 'Thought experiments', *Metaphilosophy* 36: 328-47.

Foot, Philippa. 2002. 'The problem of abortion and the doctrine of double effect', in *Virtues and Vices And Other Essays in Moral Philosophy*. Oxford: Clarendon Press, 19-32.

Gendler, Tamar Szabó. 1998. 'Galileo and the indispensability of scientific thought experiment', *British Journal for the Philosophy of Science* 49: 397-424.

Haggqvist, Sören. 1996. *Thought Experiments in Philosophy*. Stockholm: Almqvist & Wiksell International.

Hare, R. M. 1981. *Moral Thinking*. Oxford University Press.

Martin, C. B. 1958. 'Identity and exact similarity', *Analysis* 18: 83-7.

Mulhall, Stephen. 2002. 'Fearful thoughts', *The London Review of Books* vol. 24 no. 16, 16-18.

Norton, John. 1996. 'Are thought experiments just what you thought?', *Canadian Journal of Philosophy* 26: 333-66.

Nowell-Smith, P.H. 1954. *Ethics*. London: Penguin Books.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Oxford: Blackwell.

Parfit, Derek. 1986. *Reasons and Persons*. Oxford University Press.

Putnam, Hilary. 1975. 'The meaning of "meaning" ', in Keith Gunderson (ed.) *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press, 131-93.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Raz, Joseph. 1986. *The Morality of Freedom*. Oxford University Press

Rivera-Lopez, Eduardo. 2005. 'Use and misuse of examples in normative ethics', *The Journal of Value Inquiry* 39: 115-25.

Singer, Peter. 1972. 'Famine, affluence, and morality', *Philosophy and Public Affairs* 1: 229-43.

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

Thomson, Judith Jarvis. 1971. 'A defense of abortion', *Philosophy and Public Affairs* 1: 47-66.

Thomson, Judith Jarvis. 1985. 'The trolley problem', *Yale Law Journal* 94: 1395-1415.

Walzer, Michael. 1973. 'The problem of dirty hands', *Philosophy and Public Affairs* 2, 160-80.

Williams, Bernard. 1960. 'Bodily continuity and personal identity', *Analysis* 20: 117-20.

Williams, Bernard. 1973. 'A critique of utilitarianism', in J.J.C. Smart and Bernard Williams (eds.) *Utilitarianism: For and Against*. Cambridge University Press, 75-150.

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

Winch, Peter (1972), 'The Universalizability of Moral Judgments', in *The Monist …*

Wittgenstein, Ludwig (1965), 'A Lecture on Ethics', *The Philosophical Review*, 74: 3-12.