Trapped in an Experience Machine with a Famous Violinist:
Thought Experiments in Normative Theory
[July 2011]

Kimberley Brownlee and Zofia Stemplowska

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. In philosophy, some thought experiments are highly influential, even famous, such as the Trolley,[1] the Transplant,[2] Amoeba-like Persons,[3] Rawls's Original Position,[4] the Experience Machine,[5] the Utility Monster,[6] and the Ticking Bomb.[7]

However, unlike in mathematics and science, in normative theory and in philosophy generally, the use of thought experiments is a matter of controversy. Two pressing objections against their use are the following:

---

1 Where we are asked whether it is permissible to re-direct a run-away trolley from a track where it would kill five people to a track where it will kill one. Cf. Philippa Foot and JJ Thomson in Fischer, John Martin and Mark Ravizza (eds.) (1992), *Ethics: Problems and Principles*, (Orlando, Fla.: Harcourt Brace Jovanovich).

2 Where we are asked whether it is permissible to kill one person to redistribute her bodily organs to save five people. Cf. Foot and Thomson in Fischer and Mark Ravizza (eds.) (1992).

3 Where we are asked about the identity of the people who split, like amoeba, from one into two. Cf. Martin, C. B. (1958), 'Identity and Exact Similarity', *Analysis*, 18: 83-7 and Williams, Bernard (1960), 'Bodily Continuity and Personal Identity', *Analysis*, 20: 117-20.

4 Where we are asked what distributive deals would be agreed to by people deprived of certain types of knowledge. Cf. Rawls, John (1971), *A Theory of Justice*, (Cambridge, Mass.: Harvard University Press).

5 Where we are asked, for example, whether we would miss out on anything of value by leading a life inside a machine generating happy mental states. Cf. Nozick, Robert (1974), *Anarchy, State, and Utopia*, (Oxford: Blackwell), 42-45.

6 Where we are asked to consider what is owed to someone who can thrive on the suffering of others) Nozick (1974), 41.

7 Where we are asked whether it might be permissible to torture someone in order to find out a location of a bomb that is set to go off and kill many people. Cf. Walzer, Michael (1973), 'The Problem of Dirty Hands', *Philosophy and Public Affairs*, 2: 160-180, 166-167.

1. Thought experiments both invite systematic bias and entrench existing biases. (The Objection of Bias)[8]

2. Thought experiments often are inherently ambiguous, leading to inescapably opaque judgements. (The Objection of Inherent Ambiguity)[9]

These objections are troubling because they challenge the very possibility of making logically and philosophically respectable use of thought experiments. Neither objection is that 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 normative theory where thought experiments are deployed not only for conceptual and logical purposes, but also for normative and evaluative purposes. Using thought experiments in normative theory in particular may seem suspect because such thought experiments abstract away from and idealise real-life cases or even invent fantastical scenarios, but nonetheless purport to guide real-life behaviour.[10]

This paper aims to defend the use of thought experiments in 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 is to flesh out some of 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

---

8 This is a common worry. In what follows we propose 'light-touch' solutions and reject more radical ones as developed by Häggqvist, Sören (1996), *Thought Experiments in Philosophy*, (Stockholm: Almqvist & Wiksell International), 147; and Rivera-Lopez, Eduardo (2005), 'Use and Misuse of Examples in Normative Ethics' *The Journal of Value Inquiry*, 39: 115–125.

9 This worry is toyed with by Parfit, Derek (1986), *Reasons and Persons*, (Oxford University Press), 389 and endorsed by Cooper, Rachel (2005), 'Thought Experiments', *Metaphilosophy*, 36: 328-47 and Wilkes, Kathleen V. (1988), *Real People: Personal Identity without Thought Experiments*, (Oxford: Clarendon Press). See also Raz, Joseph (1986), *The Morality of Freedom*, (Oxford: Oxford University Press), 419-420.

10 The use of thought experiments in normative theory is also subject to further, less weighty objections: first, that such thought experiments are in poor taste since they often involve fantastic scenarios of suffering, death, and cruelty that trivialise that suffering, second, that they impoverish our understanding of urgent problems, as they are devoid of rich social context, and third, that thought experiments, such as the Ticking Bomb, 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 normative theory. They seem to invite theorists to engage in careful and tactful delineation of the thought experiments rather than to abandon them altogether.

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 1). We then respond to the Objection of Bias and 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 2), and second, by showing that these conditions do not preclude the use of thought experiments that involve practical impossibilities or imaginatively opaque components (Section 3).

1. Definitions

We understand thought experiments in normative theory 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.[11]

The reference here to 'mental visualisation' 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 visualisation. This does not mean that thought experiments cannot intelligibly and profitably deploy concepts that defy visualisation, 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.

---

[11] 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.

Our conception of thought experiments builds on the work of others: Thought experiments have been characterised as (1) devices 'of the imagination used to investigate the nature of things' (Brown, James Robert (2007), 'Thought Experiments' in *The Stanford Encyclopedia of Philosophy*. (Fall 2009 Edition), Edward N. Zalta (ed.)), (2) picturesque arguments (Norton, John (1996), 'Are Thought Experiments Just What You Thought?' *Canadian Journal of Philosophy*, 26: 333-66, 334), (3) purely mental procedures that aim to reveal something about the relationship between two or more variables (Sorensen, Roy A. (1992), *Thought Experiments*, (Oxford: Oxford University Press), 186 and 205), (4) judgments about what would be the case if the particular state of affairs described in some imaginary scenario were actual (Gendler Szabo, Tamar (1998), 'Galileo and the Indispensability of Scientific Thought Experiment', *The British Journal for the Philosophy of Science*, 49: 397-424, 398. Cited in Cooper, Rachel (2005), 'Thought Experiments', *Metaphilosophy*, 36: 328-47, 328-29.

1.1 Descriptive Hypothetical Examples versus Thought Experiments

The reference to 'mental visualisation' should not obscure the fact that thought experiments are only a subset of a broader category of hypothetical scenarios that involve visualisation 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 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.[12]

1.2 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.*[13] The latter are mental procedures that aim simply to predict or reveal to us our psychological/mental states. A psychological experiment asks such things as: 'Can you make yourself believe you are 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

---

12 Although we do not examine descriptive hypothetical examples here, it is worth noting two features of them in relation to thought experiments. First, descriptive hypotheticals can be proto-thought experiments that might 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. We return to this in Section 2.1a below.

13 In what follows, we shall refer to *introspective psychological experiments* simply as *psychological experiments*. Cf. Sorensen (1992), 2008-9 for discussion of 'internal psychological experiments'. For details of his rich taxonomy of experiments, see Sorensen (1992) and Sorensen, Roy A., 'Sorensen's Reply to Bunzl and Feldman', *Informal Logic* 17 (1995): 399-405.

generation of a given mental state is precisely and uniquely what is being tested.[14] 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 see how she will react). The object 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 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 (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

[14] Of course, most, if not all, imaginative mental processes are 'experiments' that generate mental states. If we are asked to imagine a Transplant Case and we go along with it, we are de facto 'experimenting' in the relevant sense in that we are triggering, first, a preliminary mental state (a visualization of the case), and then a further mental state (a belief that it is impermissible to carve up people). Our point is that sometimes finding out what these mental states are is not the object of the experiment.

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. 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 normative 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?'[15] We stress the distinction between thought experiments and psychological experiments to emphasize that thought experiments are not intended to elicit raw, unreflective intuitions or brute reactions. Their results can and often should be the fruit of reflection. And 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.

1.3 Simple Thought Experiments versus Complex Thought Experiments

Within the category of *thought experiments*, there are further conceptual distinctions. The first is between simple thought experiments and complex thought experiments. A simple thought experiment, such as the Trolley Problem, considers a single scenario. In normative 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, or good. For instance, Trolley raises the question of whether it is permissible to turn the trolley and divert the harm from the five to the one. Oftentimes, the philosopher'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 Trolley and Transplant, considers two or more scenarios in relation to each other. It contrasts (the simple thought

---

15 As noted above, 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.

experiment) Trolley with (the simple thought experiment) Transplant. It aims to establish whether our normative answers in the one case align with our answers in the other to expose a disanalogy or confirm an analogy, to undermine or affirm a hypothesis, to reveal a conflation of concepts or principles, or to bring 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.

1.4 Contingency, Necessity, and Imaginability

The final set of conceptual distinctions to highlight within the category of *thought experiments* relate to, first, their differences in degree of practical possibility and, second, their differences in degree of imaginative clarity.

Thought experiments take forms of greater or lesser practical possibility. 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 former can be described as *contingently hypothetical* (e.g. 'Imagine that I have picked up the cup in front of me.' or 'Imagine that you are walking by a pond and you spot a drowning child.'). 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, every-day reality. Being necessarily hypothetical in either of the two senses just noted is then 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 involves an imaginatively opaque being whose pleasure in sacrificing others must be of a fantastic quality so that it can outweigh the suffering of those sacrificed. What is imaginatively opaque to us as ordinary

creatures with ordinary abilities for happiness is what such fantastic happiness would involve.[16]

These two senses (and sources) of wackiness can overlap but are conceptually distinct since some cases of practical impossibility, such as your jumping 100 feet in the air, are nonetheless readily imaginable and some cases of imaginative opacity, such as the experience of being a bat or of sleepwalking, or of insanity, are nonetheless readily practically possible and indeed actual. Commonly cited examples of wacky thought experiments include Rawls's Original Position, the Experience Machine, and Amoeba-like Persons. Since imaginatively opaque and necessarily hypothetical thought experiments invite the most controversy in normative theory, they will be the main focus of our defence in Section 3.

2. Necessary Conditions for Well-Posed Thought Experiments

What would a well-posed thought experiment in normative theory look like? In this section, we outline and defend two necessary conditions for well-posed thought experiments in normative theory: 1) philosophical respectability, and 2) argumentative relevance. In broad terms, these conditions of well-posedness extend to thought experiments in domains other than normative theory, but they are particularly salient to normative theory given the controversy over the use of thought experiments for normative purposes. 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.

2.1. Philosophical Respectability

This condition has two distinct dimensions, the first of which is non-question-beggingness, which applies straightforwardly to both simple and complex thought experiments. Thought experiments should not assume an answer to the question that they pose. So, when

16 Parfit, Derek (1986), *Reasons and Persons*, (Oxford University Press), 389.

formulating thought experiments, one cannot assume that persons could divide in an amoeba-like fashion in order to argue that persons can divide in such a fashion.

The second dimension – validity – applies to all complex thought experiments and to some simple ones (as we explain below). 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 a further 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 are, whatever those standards may be. Inability to translate such thought experiments into valid arguments 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 adapted from Peter Singer.

> **The Pool and the Envelope**: Imagine that you wake up one morning and from your 20th floor apartment see a child drowning in a pool that belongs to your neighbours. You can easily save the child by pressing a button that will drain the

pool. 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 pool case, must you also save the child(ren) in the envelope case?[17]

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 pool.
P2: We have a duty to save the child in the pool.
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 this 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 Pool 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

[17] Singer, Peter (1972), 'Famine, Affluence, and Morality', *Philosophy and Public Affairs*, 1: 231-2.

> 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 and not corrupt our reflections. This is addressed by the second condition, the condition of argumentative relevance.

2.2 Argumentative Relevance

Thought experiments should be designed in such a way that we can focus upon the relevant aspects of the scenario under consideration. Why this is so should be clear; 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; P* misidentifies the principle to be derived from considering the case of the sibling. Similarly, looking at a simple thought experiment, if, in Transplant, 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, amongst other things, is whether there is a normatively significant difference between killing and letting die.[18] 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

[18] For a similar point see Rivera-Lopez, Eduardo (2005), 'Use and Misuse of Examples in Normative Ethics' *The Journal of Value Inquiry*, 39: 115–125.

support or challenge the argumentative move at issue. 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.

2.2a Rudimentary Alternatives Constraint

Concerning the Rudimentary Alternatives Constraint, when we assume the absence of some (believed) necessary feature of the world, we should stipulate an at least rudimentary alternative. The aim here is to eliminate the bias in our analysis of thought experiments 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 appropriately 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 appropriately 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.[19] Unlike the Rudimentary Alternatives Constraint, the Fleshed-Out Alternatives Constraint would require the ancient philosopher to explain *how* a large, heavy rock can be pushed by a gust of wind. We 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 need a clear, fleshed-out statement of the alternative to the ruled-out feature of the scenario in order to prevent the ruled-out feature from determining our conclusions. For example, the Fleshed-Out Alternatives Constraint would make it (implausibly) the case that no one could *entertain*, say, the certainly widespread idea that Jesus was the son of God rather than the son of Joseph and extrapolate from that.

All in all, then, we accept that implicit 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.

2.2b Moderate Narrative-Framing Constraint

Turning to narrative bias, 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, 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 an example that we owe to Roy Sorensen.[20] 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

---

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

20 Sorensen (1992), 264. The original thought experiment involving personal identity and teleportation is due to Parfit (1986), 199–200. For a discussion see, for example, Coleman, Stephen (2000), 'Thought Experiments and Personal Identity', *Philosophical Studies*, 98: 53–69, 58-60.

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 this scenario be irretrievably boring. We can rewrite this kind of scenario as an exciting, Matrix-style adventure that eliminates the supposed anti-boredom bias. Our rewriting 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, 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*.[21] 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

[21] Proposed by Haggqvist, S. (1996), *Thought Experiments in Philosophy*, (Stockholm: Almqvist & Wiksell International), 147; and developed by Rivera-Lopez (2005).

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 is to spot certain linguistic patterns in English, or a person unfamiliar with a given family's dynamics might be quicker than members are to spot 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.

All 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 mother or a saintly Mafioso might be hard to execute. Similarly, we might also worry in relation to, say, Ticking-Bomb that it asks us to assume what is hard to imagine, namely that 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.

3. 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 we are not moved by suggestions to bar wackiness. The 'wackiness' of thought experiments can be disambiguated into the two main categories noted in the Introduction: (i) imaginative opacity, and (ii) 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.

3.1. Imaginative Opacity

Turning to imaginative opacity, this dimension of wackiness raises the following worry. 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 behavior of the entity she is imagining. Or she may have knowledge of the laws relevant to predicting that behavior 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. 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, such as 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 paper) 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 do to play a given part and react 'in-character'.

3.2. 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.[22] 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, people do not split like amoebas; dogs do not speak; 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*.

Why can we apparently not know what to say about this world? The answer relates to 'semantic holism'.[23] The idea is that *our* concepts developed to track *our* world, rather than the wacky worlds that we set up in our experiments, and the latter defy plausible description with our real-life concepts. 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

---

22 The worry is developed by Nowell-Smith, P.H. (1956), *Ethics*, (London: Penguin Books). See also Raz, Joseph (1986), *The Morality of Freedom*, (Oxford: Oxford University Press), 419-420; Mulhall, Stephen (2002), 'Fearful Thoughts', *The London Review of Books*, 24 (16), 16-18.

23 For further discussion and rejection of 'semantic holism' see Sorensen (1992), 282-284.

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, consider the paint analogy. The thought 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. Similarly, in Rich and Superrich, we examine the value of equality by looking at it under the 'light' of the foreign context in which deprivation is not at issue. The question is: Do 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. Imagine a thinker in the medieval period asking whether what makes the ruler legitimate is that God has ordained him. The medieval thought experimenter might devise the following experiment:

> **The Ruler:** Suppose that a ruler were obeyed and were considered legitimate by his people even though God had not ordained him and the people did not think that God had ordained him. Would the ruler actually be legitimate?

Assume that this medieval experimenter on reflection concludes that, no, such a ruler would be illegitimate. It would be irrelevant then to point out to the experimenter that surely the hypothesised society has a different conception of *legitimacy* since the ruler is obeyed even though no one believes God ordained him. Such an observation, though factually correct, would not change the fact that the thought experiment reveals to the experimenter that his own conception of *legitimacy* requires that the ruler be ordained by God because God, not obedience, is the source of legitimacy.

Similarly, returning to Rich and Superrich, we see that, in running this thought experiment, we do not ask what we would think if we were in such a privileged society, but rather whether we consider the inequality present in that society too unfair by our own,

current standards.[24] Pointing out that in the 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.

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 hypothesised worlds.

[24] 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 characterises 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.