

A Model for Thought Experiments

SÖREN HÄGGQVIST
Stockholm University
Stockholm, SWEDEN

I Introduction

Philosophical interest in thought experiments has grown over the last couple of decades. Several positions have emerged, defined largely by their differing responses to a perceived epistemological challenge: how do thought experiments yield justified belief revision, even in science, when they provide no new empirical data? Attitudes towards this supposed explanandum differ. Many philosophers accept that it poses a genuine puzzle and hence seek to provide a substantive explanation.¹

-
- 1 Thus James Robert Brown argues that thought experiments show empiricism to be false, since their epistemic potency can only be explained by appeal to a Platonist epistemology, while Nancy Nersessian, Nenad Miscevic, and Michael Bishop echo Mach in holding that their epistemic relevance comes from the fact that they are simulations run on mental models of the world. See J. R. Brown, *The Laboratory of the Mind* (London: Routledge 1991); J. R. Brown, 'Thought Experiments: A Platonic Account,' in *Thought Experiments in Science and Philosophy*, T. Horowitz and G. Massey, eds. (Savage, MD: Rowman & Littlefield 1991); J. R. Brown, 'Why Thought Experiments Transcend Empiricism,' in *Contemporary Debates in Philosophy of Science*, C. Hitchcock, ed. (Oxford: Blackwell 2004); J. R. Brown, 'Peeking Into Plato's Heaven,' *Philosophy of Science* 71 (2004) 1126-38; J. R. Brown, 'Thought Experiments in Science, Philosophy, and Mathematics,' *Croatian Journal of Philosophy* 19 (2007) 3-27; N. Nersessian, 'How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science,' in *Cognitive Models of Science*, R. Giere, ed. (Minneapolis: University of Minnesota Press 1992); N. Nersessian, 'Thought

Others reject or deflate the epistemic claims made for thought experiments.²

In this paper I present a model for many thought experiments in philosophy and science.³ The model doesn't assume that thought experiments in fact manage to achieve epistemic justification, but it allows us to see how they aspire to do so. It also emphasises both the parallels and the discrepancies between thought experiments and ordinary scientific experiments. And it indicates that there is a systematic mismatch

Experimenting as Mental Modeling,' *PSA 1992, vol.2* 291-301 (East Lansing, MI: Philosophy of Science Association 1993); N. Miscevic, 'Mental models and thought experiments,' *International Studies in the Philosophy of Science* 6 (1992): 215-26; M. Bishop, 'An Epistemological Role for Thought Experiments,' in *Idealization IX: Idealization in Contemporary Physics, Poznan Studies in the Philosophy of the Sciences and the Humanities, vol. 63*, Niall Shanks, ed. (Amsterdam: Rodopi 1998). Tamar Szabó Gendler, who develops Kuhn's idea that thought experiments work by exposing latent conflicts in the application criteria for concepts, and Kuhn himself, also belong in this broad camp, as do the more eclectic positions of Roy Sorensen and Tim De Mey. See T. S. Gendler, 'Galileo and the Indispensability of Scientific Thought Experiment,' *British Journal for the Philosophy of Science* 49 (1998) 397-424; T. Gendler, 'Thought Experiments Rethought — and Reperceived,' *Philosophy of Science* 71 (2004) 1152-63; R. Sorensen, *Thought Experiments* (Oxford: OUP 1992); T. S. Kuhn, 'A Function for Thought Experiments,' in *The Essential Tension* (Chicago: Chicago University Press 1977); T. De Mey, 'Thinking Through Thought Experiments' (PhD dissertation, University of Ghent 2003). A position according to which thought experiments may be fruitful, but only under tightly circumscribed historical conditions, is advocated by James McAllister. See J. McAllister, 'The Evidential Significance of Thought Experiment in Science' *Studies in the History and Philosophy of Science* 27 (1996) 233-50.

- 2 Thus Kathleen Wilkes argues that thought experiments in philosophy of mind typically fail to provide good evidence. See K. Wilkes, *Real People. Personal Identity without Thought Experiments* (Oxford: OUP 1988). And John Norton claims that thought experiments are simply arguments, albeit dressed up in exotic narrative garb. See J. Norton, 'Thought Experiments in Einstein's Work,' in *Thought Experiments in Science and Philosophy*, T. Horowitz and G. Massey, eds. (Savage, MD: Rowman & Littlefield 1991); and 'Are Thought Experiments Just What You Thought?' *Canadian Journal of Philosophy* 26 (1996) 333-66; J. Norton, 'Why Thought Experiments do not Transcend Empiricism,' in *Contemporary Debates in Philosophy of Science*, C. Hitchcock, ed. (Oxford: Blackwell 2004); J. Norton, 'On Thought Experiments: Is There More to the Argument?' *Philosophy of Science* 71 (2004) 1139-51. Adherents of the argument view also include Nicholas Rescher, Andrew Irwine and John Forge. See N. Rescher, 'Thought Experimentation in Presocratic Philosophy'; A. Irwine, 'On the Nature of Thought Experiments in Scientific Reasoning'; and J. Forge, 'Thought Experiments in the Philosophy of Physical Science,' all in *Thought Experiments in Science and Philosophy*, T. Horowitz and G. Massey, eds. (Savage, MD: Rowman & Littlefield 1991).
- 3 Some arguments in this paper parts draw partly on my *Thought Experiments in Philosophy* (Stockholm: Almqvist & Wiksell International 1996).

between the epistemic pretensions of many thought experiments and what they deliver. So although the alleged epistemic merits of thought experiments is not the principal focus of this paper, the view of them that I propose bears on that issue.

The term ‘thought experiment’ is quite vague. What I am interested in, and shall denote by the term, are hypothetical cases intended to function as experiments, in the following sense: they aspire to *test* hypotheses or theories.⁴ So I will not discuss hypothetical cases intended merely to illustrate or convey a theory, although many of these are plausibly called ‘thought experiments’ too.

The outline of the paper is as follows. In the next section, I defend my stipulative restriction to hypothetical examples functioning as test cases. Section III argues that although thought experiments are not identical to arguments, they have to be seen as intimately connected to certain arguments. In section IV, I make a suggestion concerning the form of such arguments. Section V illustrates how the schema works by applying it to two influential thought experiments, Putnam’s Twin Earth case and Einstein’s clock-in-a-box thought experiment. The final section draws some morals concerning the epistemic value of thought experiments in general.

To some, this paper may seem under-ambitious in that it doesn’t head-on address the issues where most of the action is in current work on thought experiments: their epistemic merits and their underlying psychological implementation.⁵ However, I believe that having a — somewhat abstract — framework for thinking about what thought experiments aim at and how they work is both a useful propaedeutic to discussing questions of merit or implementation, and interesting in its own right. It is also useful for thinking about the dialectical situation regarding particular thought experiments. And as will become clear in the final section, the model itself does offer some insights about the reliability of thought experiments.

II What are these things called thought experiments?

It may seem that anyone setting out to make general claims about the workings of thought experiments ought to start by considering what they are. If this is taken to involve a general characterisation of the

4 Not everything called ‘experimentation’ meets this condition, of course; the term may refer simply to someone’s doing something to see what happens.

5 One of my referees in effect articulated this complaint.

plethora of things that have on some occasion been called ‘thought experiments,’ however, the task seems both daunting and misguided. For these are very different things indeed, ranging from mathematical arguments, pre-Socratic reasoning⁶ and Husserlian eidetic variation⁷ to Harvey’s discovery of the circulation of the blood.⁸ It seems quite obvious that the class of things to which the term ‘thought experiment’ has been applied does not constitute any natural kind or category.

Hence I am also skeptical of James Robert Brown’s optimism when he states that we may rely on our intuitive ability to correctly sort cases into the extension and counterextension of the term, and therefore need not bother with defining it. Brown writes: ‘We know them when we see them, and that’s enough to make talking about them possible.’⁹ He goes on to argue that we shouldn’t build into our conception of thought experiments e.g. that they involve idealization or be essentially non-executable as physical experiments: ‘These are things to be argued, debated, and, with luck, resolved at the end of inquiry, not fixed by stipulation at the outset.’¹⁰ However, if thought experiments aren’t a natural kind but rather the opposite, a bunch of highly heterogeneous things, then it is hard to see how there can be either inquiry or meaningful debate before the participants have agreed on some more restricted understanding of the term. Trying to articulate the character or workings of all the things called thought experiments seems likely to result in a ‘theory’ of an overly eclectic and gerrymandered sort.

The most attractive option, therefore, seems to be to stipulate what we are to talk about. I believe that a restriction to hypothetical examples used as tests of theories captures a class of things worth theorising about. I also think it is inclusive enough to harbour almost all exemplars standardly alluded to in the literature on thought experiments: Galileo’s experiment on falling bodies, Newton’s thought experiments aimed to show the existence of absolute space, Einstein’s, Podolsky’s, and Rosen’s objection to the Copenhagen interpretation of quantum

6 See N. Rescher, ‘Thought Experimentation in Presocratic Thinking,’ in Horowitz and Massey.

7 See J. N. Mohaney, ‘Method of Imaginative Variation in Phenomenology,’ in Horowitz and Massey. The point that the term encompasses very diverse things was forcefully made by Tamar Gendler in a review of Horowitz and Massey; see T. Gendler, ‘Tools of the Trade: Thought Experiments Examined,’ *The Harvard Review of Philosophy* VOL?? (1994) 81-5.

8 See De Mey, ‘Thinking Through Thought Experiments.’

9 ‘Why Thought Experiments Transcend Empiricism,’ 25

10 *Ibid.*, 26

mechanics, Searle's Chinese Room, Jackson's Knowledge argument, Putnam's Twin Earth; etc. To help the reader accept this claim, two minor clarifications may be in order. First, 'hypothetical' doesn't entail non-actuality, only that the situation contemplated in the thought experiment is entertained as a possibility in thought. Second, 'test' doesn't entail any particular epistemic attitude on the thought experimenter's part towards what is tested — certainly not that she be agnostic about it prior to the execution of the thought experiment.¹¹

According to Brown, the most interesting thought experiments (which he calls 'Platonic') are those which not only refute an existing theory but also suggest a new one which they lend rational support. It may fairly be asked whether my stipulation doesn't unduly rule out consideration of these cases.¹²

But consider: the only Platonic thought experiments mentioned by Brown are just three: EPR against the Copenhagen interpretation of quantum mechanics, Galileo's on falling bodies, and Leibniz' on *vis viva*. Of these, the third is basically a calculation which reduces to absurdity Descartes' theory that conserved force is (something like) mass \times velocity and shows that it is quite generally mv^2 (which Leibniz called *vis viva*). The generality of the positive conclusion really has nothing to do with contemplation of a hypothetical scenario. One may call this a thought experiment if one likes, but not accommodating it seems a small price to pay compared to the gains achieved by the restriction I suggest.

As for the other two exemplars of Platonic thought experiments adduced by Brown, their negative or destructive aspects are most plausibly construed as purported modal counterexamples to Aristotle's theory on falling bodies and the Copenhagen interpretation, respectively. And as I have argued elsewhere, the 'new theories' that Brown claims are suggested and given credence by them amount to little more than the negations of the theories they seek to reject.¹³ Hence my claim

11 I shall argue below that thought experiments are typically devised in order to refute a theory one already takes to be false (just like many ordinary experiments), and to induce others to share this belief.

12 An anonymous referee did just that.

13 See Häggqvist, *Thought Experiments in Philosophy*. In the case of Galileo, the thought experiment has to be buttressed by a thought experiment refuting the reverse of Aristotle's actual theory, i.e. the claim that light bodies fall faster than heavy ones. Then the theory suggested by the thought experiment — that bodies fall at the same speed regardless of weight — emerges as the negation of the disjunction of the two refuted theories. For more on Brown's treatment of Galilei, as well as an extended discussion of his views on thought experiments, see S. Häggqvist, 'The

above that the restriction admits inclusion of these admittedly interesting cases.

Finally, let me stress that the stipulative restriction I suggest is quite often taken to be a natural precisification of the term. For instance, in her contribution to a recent PSA symposium on thought experiments, Tamar Szabó Gendler writes: 'I will assume that to perform a *thought experiment* is to reason about an imaginary scenario with the aim of confirming or disconfirming some hypothesis or theory'¹⁴ And Tim Williamson talks of 'the use of imaginary counterexamples supposedly to refute philosophical analyses or theories,'¹⁵ a practice discussed in his recent book under the rubric 'Thought Experiments.'¹⁶

III Thought experiments and arguments

If one accepts the claim that thought experiments do indeed furnish epistemic justification, at least sometimes, but rejects the claim that they do so through a special faculty of non-empirical quasi-perception, it may seem reasonable to insist that thought experiments are simply a breed of arguments. Several philosophers have done just that. John Norton holds that thought experiments are 'arguments which: (i) posit hypothetical or counterfactual states of affairs, and (ii) invoke particulars irrelevant to the generality of the conclusion.'¹⁷ Norton also suggests that this is the sole alternative to Brown's Platonism: 'The alternative to this view is to suppose that thought experiments provide some new and even mysterious route to knowledge of the physical world.'¹⁸ The argument view is appealing not least since it offers a straightforward answer to the question of *how* thought experiments may provide fresh insight: this is what inference does all the time. Moreover, its naturalness is bolstered by the widespread practice of *calling* well-known thought

A Priori Thesis: A Critical Assessment,' *Croatian Journal of Philosophy* 19 (2007) 47-61. For instructions on how to run the reverse experiment, see De Mey, 'Thinking Through Thought Experiments.'

14 'Thought Experiments Rethought — and Reperceived,' 1154. Although they may not entirely endorse this conception, none of her three co-symposiasts — James McAllister, James Robert Brown and John Norton — actively opposed it.

15 See T. Williamson, 'Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking,' *Proceedings of the Aristotelian Society* 105 (2005) 1-23.

16 Chapter 6 of T. Williamson, *The Philosophy of Philosophy* (Oxford: Blackwell 2007).

17 'Thought Experiments in Einstein's Work,' 129.

18 *Ibid.*

experiments 'arguments.' Thus Searle's Chinese Room thought experiment is frequently referred to as 'Searle's Chinese Room argument'; we have 'Putnam's Twin Earth argument,' 'Newton's bucket argument,' Jackson's 'Knowledge argument,' and so on.

Nevertheless, the argument view is not quite correct. There are various ways of seeing this. One is to notice, as Michael Bishop does, that whereas instances of the same argument type must have identical conclusions, different conclusions may be drawn from the same thought experiment on different occasions and by different thinkers.¹⁹ Another is to stress, with Roy Sorensen, the parallels between thought experiments and ordinary experiments, and to argue that since the latter are not plausibly classified as arguments, the former aren't either.²⁰ In view of the stipulation that 'thought experiments' is to refer to things intended to function as experiments, such symmetry seems both mandatory and reasonable.

In my view, however, there are additional considerations telling against the argument view. Perhaps the most striking difference between arguments and thought experiments is that unlike arguments, thought experiments are not composed of truth-valued entities. Nor are they valid or invalid in any formal sense (although they may be more or less felicitous or successful).²¹ Just like ordinary experiments, they are plausibly held to be (types of) processes, events or procedures. Their location is different, of course: the processes that constitute (a token of) a physical experiment take place in the laboratory (or like); the processes that constitute (a token of) a thought experiment are psychological and take place largely inside the thought experimenter's skull, although they may also involve such prosthetic devices as pen and paper or computers.²²

19 M. Bishop, 'Why Thought Experiments Are Not Arguments,' *Philosophy of Science* 66 (1999) 534-41.

20 R. Sorensen, *Thought Experiments* (Oxford: OUP Press 1992), 214.

21 Of course, Norton would disagree with this (as a referee remarked), so in a sense this claim begs the question against Norton. But I submit that most theorists not already wedded to the argument view will find this observation plausible. And Sorensen's and Bishop's points speak independently against the argument view.

22 After all, some thinking arguably takes place outside our bodies in such parts of what, following Dawkins, might be considered our extended phenotype. See R. Dawkins, *The Extended Phenotype* (San Francisco: Freeman 1982). A corollary to viewing thought experiments as psychological processes is that they are not properly contrasted with *real* experiments, since such processes are themselves patently real.

Hence, it seems most reasonable not to identify thought experiments with arguments. Yet I think that the argument view captures important insights worth salvaging.

For one thing, both experiments and thought experiments work only through their connection with arguments. There would be little point in conducting experiments and observing their outcomes unless these procedures had some bearing on the truth-value of theories or hypotheses. But we assess the truth-value of *theories* (as opposed to observation statements) by appeal to arguments. If experiments are not arguments, and themselves lack truth-value, what is the connection? I suggest that it is causal and, in broad stroke, this: events taking place in a laboratory experiment causes observers in the lab to hold certain observation statements true, which may then be employed in arguments concerning the theory to be tested. And performing a thought experiment causes thought experimenters — inventor or audience — to hold certain non-observational statements true, which may subsequently be employed in arguments concerning the theory to be tested. Thus both experiments and thought experiments are conducted to generate (acceptance of) premises for certain arguments concerning some theory or hypothesis.²³

If this suggestion is correct, the practice of calling thought experiments arguments is readily explained by the intimacy of the causal connection, together with the fact that thought experiments are conveyed in (and reproduced by means of) linguistic form, thereby inviting a construal of them as arguments. But it raises other issues. One pressing question clearly is whether the causal processes leading from experiment to belief formation are reliable. Although I will not try to assess their reliability in the case of experimentation in general, I shall argue that they are bound to be less reliable in the case of thought experiments than in the case of ordinary experiments. But another, and prior, question is what sorts of arguments stand in such connection to thought experiments.

IV Thought experiments regimented

There are obviously several different arguments with which even a single thought experiment may be connected, in the sense described above. It would be nice to have an argument schema that is general enough to

23 Again, this is not the only use of everything reasonably called 'thought experiments.'

permit systematic discussion of many different thought experiments, as well as responses to them, yet fine-grained enough to permit at least some recognition of the particular thought experiment it is connected to. I should also like to make explicit the fact that a thought experiment is typically designed to invite the conclusion that the target thesis is false (why this is so will be touched upon briefly later in this section). I propose the following argument schema, where 'C' is the counterfactual scenario described in the thought experiment, 'T' is the theory to be tested, and 'W' is a statement claimed by the thought experimenter to be (i) false in the counterfactual scenario, yet (ii) one to whose truth the theory under testing is committed in that scenario.²⁴

$$\begin{array}{l}
 (\alpha) \quad \diamond C \\
 \quad T \supset (C \Box \rightarrow W) \\
 \quad C \Box \rightarrow \neg W \\
 \hline
 \quad \neg T
 \end{array}$$

Arguments instantiating this schema will be valid, on e.g. a Lewis semantics for counterfactuals, since the first and the third premise jointly imply $\neg(C \Box \rightarrow W)$.

One benefit of this schema is that it does seem to reflect the dialectical progression of most thought experiments. First, a counterfactual scenario presented as (in some sense) possible is offered for contemplation; then the thesis under attack is held to be connected to a certain consequence in that scenario; whereupon the non-emergence of that consequence in that scenario is asserted.

Another benefit is that the schema parallels a schema for arguments connected to ordinary experiments. When such an experiment leads to revision or rejection of a previously entertained theory or hypothesis (T), it may be said to do so by way of a conditional ($I \supset O$) predicting a certain outcome (O) given certain initial conditions (I). This conditional is connected to T via a nested conditional $T \supset (I \supset O)$; the falsity of T is inferred from this along with a premise saying that I holds and another premise saying that O is false. Schematically:

24 As a matter of etiology, this schema emerged from reflections on Sorensen's suggestions for regimentation in *Thought Experiments*, ch. 6. But as a referee pointed out, it seems a pretty straightforward proposal once one thinks of thought experiments as aiming at testing theories. For discussion of what I take to be the shortcomings of Sorensen's proposal, see my *Thought Experiments in Philosophy*, ch. 5.

(A) $T \supset (I \supset O)$

I

 $\neg O$

 $\neg T.$

The schema I have proposed for thought experiments is just a modalized version of this schema, in keeping with the initial limitation of the extension of 'thought experiment' to things intended to function as experiments.

Clearly, several different arguments of varying logical form may said to be 'connected' or 'associated' with any given thought experiment, as well as with any given experiment. The selling point of schema (α) is not uniqueness, but rather its feasibility: that it offers a neat format for discussing common important features of thought experiments and debates surrounding them; and generality: that it does in fact seem quite natural for very many thought experiments, both in philosophy and in science. These virtues will be briefly illustrated in the next section.

The model may seem vaguely Popperian in its emphasis on *counterinstances* to the target thesis. But it is a fact that this is what thought experimenters are typically in the business of trying to construct. And this is, in turn, only to be expected. Confirmation is a tricky notion in the best of settings, and in a modal setting, we really don't have to be Popperians to appreciate the impotence of positive but counterfactual instances. A thought experiment showing only that some theory would give the right result in a certain merely possible situation wouldn't pack much epistemic punch. Hence the negative slant, noted also by Sorensen,²⁵ to most epistemically ambitious thought experimentation.

With the schema (α) at hand, it is time to qualify my claim above that the connection between a thought experiment, considered as a cognitive process, and the argument which may make it useful for assessing theories is causal. The qualification I have in mind doesn't concern the question whether the experiment yields reliable belief in the premises of an argument of the form (α); this question will be broached at the end of the paper. It is just that although the thought experiment-as-process is plausibly taken to cause belief in premises of the form $\diamond C$ and $C \square \rightarrow \neg W$, it isn't plausible to hold that it causes belief in premises of the form $T \supset (C \square \rightarrow W)$. Rather, the thought experimenter tends to accept this claim antecedently, and design her example accordingly (for

25 Sorensen, *Thought Experiments*, 135

instance, in choosing a scenario she takes to be relevant to T). Again, the situation seems parallel with ordinary experiments, where the belief that a given theory implies certain conditionals is a prior, guiding assumption in the choice of what experiment to perform. However, the status of the nested conditional is peculiar in the case of thought experiments and results in problems peculiar to them, as I will explain in the final section. For the meantime, let's just record that belief in the relevant nested conditional is not *engendered* by the experiment or thought experiment itself.²⁶

Clearly, we should like our model to allow for *failed* thought experiments. And we should be able to make sense of the fact that the same thought experiment may lead thinkers to different conclusions. The model I propose meets both these desiderata. Let us start with the second.

As Bishop notes, the view that thought experiments are identical to arguments implausibly has to attribute contemplation of different thought experiments whenever two thinkers disagree about the conclusion of what seems to be one and the same thought experiment.²⁷ A view which does not identify thought experiments with arguments may escape this problem. But on my view, each thought experiment is designed to result in acceptance of a certain argument, with which it is thus connected. And diverging conclusions entails distinctness of arguments. So how does my view escape it? By not holding that there is just one argument with which a thought experiment is connected and in particular, by denying that it is only connected with arguments of form (α) . This schema corresponds to the thought experimenter's intentions — to her purposes in designing the thought experiment. But these designer intentions do not confer any epistemic privilege on her when it comes to the evaluation of the thought experiment.

On my view a thought experiment is potentially connected with four different arguments, each incompatible with the other three. Like all arguments involved in refutation, (α) may be viewed as deriving from an inconsistency, in this case the inconsistency of the set $\{T, \Diamond C, T \supset (C \Box \rightarrow W), C \Box \rightarrow \neg W\}$. Resolving the inconsistency by rejecting T corresponds to the aim of the thought experiment, hence to (α) . But there are three other minimal ways of resolving the inconsistency while saving T, corresponding to three different and, like (α) , valid argument schemas.²⁸

26 Thanks to an anonymous referee who prompted this paragraph.

27 'Why Thought Experiments Are Not Arguments'

28 The formal availability of these 'ways out' is quite parallel to the holism of hypoth-

One such way of defending a theoretical claim against an adverse thought experiment is to reject the thought experimenter's claim about what would be the case in the scenario depicted in the thought experiment. Such a defence corresponds to an argument of the form:

$$\begin{array}{l}
 (\beta) \quad T \\
 \quad \Diamond C \\
 \quad T \supset (C \Box \rightarrow W) \\
 \hline
 \quad \Diamond C \ \& \ C \Box \rightarrow W
 \end{array}$$

whose conclusion implies $\neg(C \Box \rightarrow \neg W)$. One may think of this way of defending T as *biting the bullet*. The defence denies that there is anything wrong with holding that W would be true in C; if this consequence is felt to be odd, its weirdness is perhaps blamed on the oddity (as opposed to impossibility) of the scenario.²⁹

Another reaction may be called the *irrelevance defence* and consists in denying that T is committed to W's being true in the scenario. This yields an argument of the form

$$\begin{array}{l}
 (\gamma) \quad T \\
 \quad \Diamond C \\
 \quad C \Box \rightarrow \neg W \\
 \hline
 \quad \neg(T \supset (C \Box \rightarrow W)).
 \end{array}$$

This move is sometimes made by philosophers professing to be generally skeptical of the use of thought experiments.³⁰ It may also be the

esis testing in general. For explicit discussion of the Duhemian holism in connection with thought experiments, see Häggqvist, *Thought Experiments in Philosophy* (ch. 6), and A. Bokulich, 'Rethinking Thought Experiments,' *Perspectives on Science* 9 (2001) 285-307.

29 Hardline utilitarians can sometimes be observed making this move.

30 E.g. Davidson: 'I have a general distrust of thought experiments that pretend to reveal what we would say under conditions that in fact never arise.' ('Epistemology Externalized,' reprinted in D. Davidson, *Subjective, Intersubjective, Objective* [Oxford: OUP Press 2001]). Davidson writes this a propos Burge's and Putnam's thought experiments in favour of externalism; for the record, let's note that the comment inaccurately characterizes the intent of these cases: their authors pronounce on what we in the actual world *should* say *about* the counterfactual cir-

rationale behind lay people's dismissal of counterfactual questions as 'merely hypothetical'.³¹ But sometimes careful argument is adduced to show that although the scenario is possible and the untoward statement instantiating W would indeed be false if it were true, T is compatible with this.³² In effect, this reply tends to claim that although the scenario is possible and the thought experimenter right about what would hold in it, it is too far out in logical space to be relevant to the theory under attack.³³ But it may also be grounded in a claim that the theory, understood aright, doesn't commit itself to the false claim about the scenario, independently of such considerations about modal distance.

The fourth type of response might be called the *impossibility defence* and denounces the scenario as impossible, thereby allowing both the thought experimenter's claim about what would hold in the scenario and the contrary counterfactual to be true at the same time. This gives an argument of the form

$$\begin{array}{l}
 (\delta) \quad T \\
 \quad T \supset (C \Box \rightarrow W) \\
 \quad C \Box \rightarrow \neg W \\
 \hline
 \quad \neg \Diamond C.
 \end{array}$$

Clearly, these various responses will not tend to be equally plausible. But they are always formally available. And interestingly, there are many cases where moves corresponding to these responses are actually made in the debates surrounding thought experiments. The model thus offers a classification, albeit coarse, of these moves. In this way, I suggest, it is useful as a tool for understanding the dialectical situation in those debates.

Although quite common, arguing along the lines of (β) , (γ) , or (δ) is clearly not the only way to reject a thought experiment. For one may

cumstances they depict (and hence what would be the case in them), not what we would say *in* them.

31 For a discussion of such hypophobia, see Sorensen, *Thought Experiments*, 275.

32 In section VI below, I argue that this response is harder to assess than the corresponding response in the case of an ordinary experiment, and that this is detrimental to the epistemic value of thought experiments.

33 This seems to be the gist of Daniel Dennett's reply to various zombie, blockhead and swampman cases. See e.g. D. Dennett, 'Get Real,' *Philosophical Topics* 22 (1994) 505-68.

hold that its target thesis is false without accepting the argument instantiating (α). In other words, one may reject more than one member of $\{T, \Diamond C, T \supset (C \Box \rightarrow W), C \Box \rightarrow \neg W\}$.

Now for failure. Here is a suggestion. A *failed* thought experiment is one whose regimentation as an instance of (α) is such that its premises are not all justified.³⁴ Whenever the premises of instances of (β), (γ), or (δ) are better justified than the premises of (α), we have a special case of this condition. Correspondingly, success may be defined as the case where the premises of a regimentation with the form (α) are all justified. Justification being relative to occasion and epistemic situation, a thought experiment may be failed on one occasion but successful on another, which is as it should be.

However, I think that pursuing the parallel with ordinary experiments may motivate a different, and stronger, notion of success, on which all, or almost all, thought experiments will be failed. This suggestion will be briefly developed in section VI.

V The model at work

Before that, let's see how the model works by applying it to a couple of famous specimens.

Putnam's Twin Earth may be regimented as the following argument.

- P1 It is possible that we had Twin Earth Doppelgängers.
- P2 If psychology determines reference/extension, then if we had Twin Earth Doppelgängers, they would refer to water with 'water.'
- P3 If we had Twin Earth Doppelgängers, they would not refer to water with 'water.'
-
- P4 Therefore, psychology does not determine reference/extension.
-

34 Insisting on *soundness* of the regimentation would be asking too much. The issue here is with justification, not truth, since we are asking when a thought experiment aiming at justified belief revision achieves or fails to achieve this aim.

Putnam's own aim was to refute the thesis that psychology determines reference/extension.³⁵ Putnam thus unsurprisingly favoured the above argument, which is an instance of schema (α). But the reaction of some philosophers has been to reject (P3). If water is not necessarily H₂O, then H₂O and XYZ might be regarded simply as two chemically different kinds of water. But then the Twin Earth Doppelgängers would refer to water by 'water' (if they existed). These philosophers thus favour the argument corresponding to schema (β). A less fashionable response, commonly dismissed as a mere quibble but quite cogent in its own right, is to reject (P1) on the grounds that there could not exist neurophysiological copies of people consisting of 70% water on a waterfree planet: hence Putnam's thought-experimental supposition is incoherent. This response corresponds to schema (δ) above. A different tack, represented by Fodor, is to reject (P2) while insisting that the target hypothesis rejected in (P4), interpreted aright, is correct although the counterfactual of (P2) is not. Fodor argues, along with many other philosophers, that narrow psychology does determine extensions of terms, but not alone: it does so only relative to an external context, which differs between Earth and Twin Earth.³⁶ This corresponds to the schematic argument (γ).

To repeat, I am not suggesting that each of these responses is equally justified or plausible. What's noteworthy here is that Putnam's thought experiment leaves formal leeway for these different resolutions; that they have actually been proposed in the discussion of Twin Earth; and that they correspond to the four different argument schemas of our model.

Let us also look briefly at an example from science. Einstein's clock-in-a-box thought experiment is interesting in its own right, but also because it was famously repudiated by Bohr. It is rare in being generally viewed as a failed thought experiment whose inventor himself graciously conceded defeat. The details of the case are, like its subject matter, intricate, but the basics are accessible and will suffice for our purposes here.

In 1930 Einstein suggested a hypothetical counter-instance to Heisenberg's uncertainty principle, according to which there is a physical limit

35 This thesis is an amalgam of two theses: (i) that speaker's psychology determines the intension of her words, and (ii) that intension determines extension/reference. So giving up the target thesis may be done by giving up either of these claims. Putnam opts for relinquishing (i). Also note that 'psychology' here means 'narrow psychology' (as it used to be before the distinction between broad and narrow psychology was invented).

36 J. Fodor, *Psychosemantics* (Cambridge, MA: MIT Press 1987)

to the accuracy with which simultaneous measurement of so-called conjugate properties can be carried out. Einstein described a hypothetical device for carrying out arbitrarily exact measurements of two conjugate properties: the energy and time of exit of a photon leaving the device. The device would consist of a box from which a single photon could be emitted at a precise time by means of a shutter mechanism connected to a clock. Once the single photon had exited, the box's change of mass could be measured by a weight. From this reading, the energy of the photon could be calculated by means of the equation $E = mc^2$.³⁷ Regimented as an argument instantiating (α), the clock-in-a-box experiment looks like this:

- (E1) It is possible (in principle) that a single photon exit a box equipped with an arbitrarily exact timer and an arbitrarily sensitive spring-balance.
- (E2) If the uncertainty principle holds, then if a single photon exited a box equipped with an arbitrarily exact timer and the box were then weighed, the time and energy of its passage would not be simultaneously measurable to any degree of accuracy violating the inequality $\Delta E \times \Delta t > h$ (where h is Planck's constant / 2π).
- (E3) But if a single photon exited a box equipped with an arbitrarily exact timer and the box were then weighed, the time and energy of its passage would be simultaneously measurable to any degree of accuracy.

(E4) Hence, the uncertainty principle doesn't hold.

Bohr retorted that the device wouldn't allow arbitrarily exact measurement. The kink is that the photon's exiting the box causes it to move in a gravitational field, which according to general relativity affects the speed of the clock, thus undermining the desired exactness of the measurement of the time of exit (or else incurring a price for such exactness in the form of lesser accuracy of the mass/energy measurement). 'Consequently, a use of the apparatus as a means of accurately measuring

37 N. Bohr, 'Discussions with Einstein on Epistemological Problems in Atomic Physics,' in *Albert Einstein: Philosopher-Scientist*, P. A. Schilpp, ed. (La Salle, IL: Open Court 1949).

the energy of the photon will prevent us from controlling the moment of its escape.³⁸

In effect, then, Bohr rejects (E3) and offers an argument instantiating (β). Or alternatively, Einstein may be taken to claim that a device allowing simultaneous, arbitrarily exact, measurement of the time and energy of the photon's exit is possible, and that if a photon exited such a device, it would produce arbitrarily exact readings for both magnitudes although the uncertainty principle predicts otherwise. Then Bohr's reply would rather constitute an impossibility proof, instantiating schema (δ). But in any event, the model seems to make the structure of the thought experiment and the disagreement between Bohr and Einstein perspicuous.

As Bishop stresses, Bohr and Einstein did indeed draw different conclusions from the same thought experiment.³⁹ But the moral Bishop correctly draws from this — that the thought experiment they both executed cannot be identified with an argument — is only part of the story. The arguments advanced by Einstein and Bohr respectively were indeed different.⁴⁰ But they were closely related: both were resolutions of the same anomaly.

Although they obviously fall short of establishing any general fertility of the model, these two cases should at least begin to illustrate how it may apply to both scientific and philosophical thought experiments.⁴¹

VI Two problems

In closing, I want to mention two features of thought experiments which are brought to light by the model, and which set them apart from

38 Bohr, 'Discussions with Einstein,' 228. A referee asks whether this doesn't concede Norton's point in reply to Bishop ('Why Thought Experiments do not Transcend Empiricism,' 63-4): that Einstein and Bohr in effect contemplate different thought experiments. Though granting this would not jeopardize my general claims, I do not think this is the right way to think about the disagreement between Einstein and Bohr, or generally between people disagreeing over a thought experiment. Such disagreement will always involve some difference in opinion, which will be reflected in the premises they accept in arguments concerning the case at hand. It seems wanton to infer that they thereby contemplate different cases. The same point obviously applies to controversies surrounding ordinary experiments.

39 'Why Thought Experiments Are Not Arguments.'

40 Or rather, they did so initially, since Einstein quickly accepted Bohr's argument and abandoned his own.

41 Several other examples are given in Häggqvist, *Thought Experiments in Philosophy*.

ordinary experiments. Both features suggest that thought experiments are less reliable than ordinary experiments.

First, consider the mechanisms resulting in belief in premises of an argument instantiating (α) when a thought experiment is performed. These may draw on all sorts of cognitive resources: theoretical and other belief, memory, inference, genetically inherited modal expectations, folk physics, and so on. If Nersessian and others are right, they may involve manipulation of mental models.⁴² If Brown is right, they may include special faculties of intellectual *Schauung*.⁴³ If David Chalmers, Frank Jackson, and an earlier time slice of Stephen Yablo are right, modal claims such as the premises of a regimented thought experiment may be justified by appeal to conceivability.⁴⁴ Some may wish to appeal to a faculty of modal intuition (although the term ‘intuition’ is more often and naturally applied to the — seemingly non-inferential — modal *belief* itself that is engendered by contemplation of the thought experiment).

Regardless of the exact character of these mechanisms, however, it seems clear that they are less reliable than the mechanisms that operate in the case of ordinary experiments. If an ordinary experiment threatening a theory T does so by giving rise to an inconsistent set of the form $\{T, T \supset (I \supset O), I, \neg O\}$, as suggested above, consistency may be achieved equally well, from a purely formal point of view, by rejecting any of its four members. But the statements corresponding to the description of initial conditions and predicted but non-emergent outcome — I and $\neg O$ — will usually be more observational than the other two. Hence, the mechanisms by which the processes that constitute the experiment give rise to belief in these statements will, on the whole, be ordinary perceptual processes. And these are quite reliable, at least to judge by the degree of intersubjective agreement they achieve. By contrast, the sheer lack of such intersubjective agreement in the case of thought experiments indicates that the corresponding mechanisms, *whichever they are* (and they will not be ordinary perceptual processes), are not equally reliable. If an observational statement involved in the evaluation of an

42 Nersessian, ‘How Do Scientists Think?’

43 Brown, *The Laboratory of the Mind*, ch. 4

44 See, e.g. D. Chalmers, *The Conscious Mind* (Oxford: OUP 1996), F. Jackson, *From Metaphysics to Ethics* (Oxford: OUP 1998), and S. Yablo, ‘Is Conceivability a Guide to Possibility?’ *Philosophy and Phenomenological Research* 53 (1993) 1-42. Yablo has since criticized the use of conceivability as evidence for modal claims; see S. Yablo, ‘Textbook Kripkeanism and the Open Texture of Concepts,’ *Pacific Philosophical Quarterly* 81 (2000) 98-122.

ordinary experiment is questioned, it is fair to invite the questioner to simply see for herself (perhaps after repeating the experiment). But the analogous reply in the case of a modal statement involved in the evaluation of a thought experiment — ‘it’s obvious; just think about it yourself’ — clearly carries much less evidential and persuasive force. Hence, the statements corresponding to the description of initial conditions and outcome — C and $C \Box \rightarrow \neg W$ — tend to be more vulnerable to rejection than their counterparts in ordinary experiments, despite the formal parallels.

With ordinary experiments, then, the reliability of the processes coupling the experiment-as-process to belief in pertinent premises tempers the holism inherent in the fact that, from a formal point of view, all that any experiment gives rise to is an inconsistency which may be resolved in several ways. With thought experiments, this tempering factor seems absent. If the psychological mechanisms linking the thought-experiment-as-process to belief in pertinent premises are not reliable, believing those premises solely on the strength of those mechanisms — i.e. solely on the basis of having performed the thought experiment — will not be justified.

So with a well-designed ordinary experiment, executing it and watching what happens typically comes close to justifying belief in certain premises in an argument concerning the tested theory or hypothesis. With thought experiments, execution leaves the question of justification open to reasonable doubt; and execution of a thought experiment typically fails to induce unanimity among its performers. So if we require that a successful experiment should induce justified belief simply on the strength of its execution, it seems that thought experiments will typically not be successful.

Is this equally the case for scientific and philosophical thought experiments? Don’t the former tend to be much less controversial than the latter?⁴⁵ In the absence of any serious empirical investigation of the matter, this claim is hard to evaluate. But it seems clear that very many thought experiments in science have been contested along the lines sketched above. Thus Leibniz disagreed with Newton about what would happen to the rotating spheres in empty space, and Bohr disagreed with Einstein, Podolsky, and Rosen on whether the experimental set-up of their famous thought experiment was possible; etc. Moreover, the absence or presence of controversy is a sociological phenomenon which often has other explanations than the inherent justification or dubiousness of the thought experiment. Some thought experiments command

45 This was suggested by a referee.

near unanimity even when they shouldn't — Twin Earth is a case in point. Finally, the border between science and philosophy sometimes isn't very sharp.⁴⁶

The second disanalogy is this. In the case of an ordinary experiment giving rise to anomaly, the nested conditional $T \supset (I \supset O)$ connecting the threatened theory T to the more observational $I \supset O$ is frequently backed by something approaching a deductive argument, at least in mature sciences. Typically, that argument relies on some (or lots of) auxiliary hypotheses. So when these auxiliary hypotheses are not beyond suspicion, it may be reasonable to stick to the threatened theory by rejecting the nested conditional. But often, in a well-designed experiment, the auxiliary hypotheses have been scrupulously tested beforehand. And even when they are suspect, it is often fairly clear *which* auxiliary hypotheses need further investigation. So although there are no crucial experiments, and no refutation by experiment may be quite conclusive, we have a fairly good grasp of what it would take to hold a theory responsible to a certain outcome under certain initial conditions.

In sum, then: when we trust our auxiliary hypotheses, it will be unjustified to reject the nested conditional, since we trust our logic. And when we don't trust our auxiliary hypotheses, at least we know what it would take to justify the inference of $I \supset O$ from T , and, quite often, how to proceed to find out whether this inference is justified or not.

But with thought experiments, the status of the corresponding nested conditional, $T \supset (C \square \rightarrow W)$, is much less clear. We usually cannot *deduce* counterfactual conditionals from philosophical or scientific theories, regardless of which auxiliary hypotheses we accept.⁴⁷ A given general statement is clearly often held to imply certain counterfactuals. And this is expressed metaphorically by saying that the theory under discussion 'yields,' 'supports,' or 'sustains' them. But these metaphors are

46 Cf. the discussion below.

47 The obvious exception being necessity claims. After this paper was submitted, Tim Williamson has suggested a formalization of philosophical thought experiments according to which they offer counterfactual counterexamples to metaphysical necessity claims, with Gettier's objections to the analysis of knowledge as justified, true belief as a prime example. See e.g. Williamson, 'Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking' and *The Philosophy of Philosophy*. The quick rejoinder is that most theses targeted by thought experiments are not metaphysical (or logical) necessity claims. For discussion of Williamson's views in general, and his formal model in particular, see S. Häggqvist, 'Modal Knowledge and the Form of Thought Experiments,' forthcoming in *The A Priori and Its Role in Philosophy*, N. Kompa, C. Nimtze and C. Suhm, eds.

not replaceable by verbs expressing well-charted logical relations. In particular, the relation is not logical consequence.

We could express the relation between a theory and a counterfactual to which we hold it responsible by saying that the counterfactual's antecedent falls within the theory's *modal scope*. Some theories have wide modal scope (for instance physical theories), some have narrow scope (for instance, perhaps, some ethical theories). But if we want to trade metaphors for exactness, 'modal scope' is hardly an improvement. I doubt that there is a general theory to be had concerning what determines modal scope.

This means that the evaluation of a thought experiment is threatened with a peculiar source of controversy. With ordinary experiments, the situation may be messy, but there is usually a way to proceed. If there is disagreement about what to hold a theory responsible for, either logic or further testing of auxiliary hypotheses may help. With thought experiments, the corresponding disagreement comes to the question whether the thought experimental scenario falls within the modal scope of the theory; hence, whether what I called the irrelevance defence — corresponding to schema (γ) — is feasible. But if there is disagreement about *this*, logic won't help at all, and it is hard to see how the dispute should be rationally settled.⁴⁸

Here too, one might perhaps think that thought experiments in science are less vulnerable in this respect.⁴⁹ Can't we simply deduce counterfactuals from theories in mature sciences? I think that any impression of an important difference here is an illusion, however, stemming from the fact that physical theories are generally taken to aspire to very wide modal scope. The situation is different in biology, for instance. Were someone to challenge a biological hypothesis by means of a thought-experimental counterexample, and someone else replied that the hypothesis never aspires to cover worlds as far out in logical space as the scenario, this controversy would hardly seem to be resolvable by formal methods. Moreover, as I mentioned above, the distinction between science and philosophy simply isn't cut and dried. Is the theory

48 Cf. Norton's characterisation of thought experiments as 'arguments which ... invoke particulars irrelevant to the generality of the conclusion' ('Thought Experiments in Einstein's Work,' 129). On the view I am advocating, any experiment aspires to deliver premises for an argument invoking particulars — those of the experimental set-up — that are *relevant* to the generality of the conclusion, by standing in an appropriate relation of instantiation or counterinstance to the general statement mentioned in the conclusion. But with thought experiments, even what *counts* as relevance is moot.

49 This was suggested by an anonymous referee.

that the types of subjective character of experience are identical to types of brains states a scientific theory or a philosophical one? The question seems barren; yet this theory has been hotly debated by means of thought experiments such as Jackson's Knowledge argument. And the reply to such examples has not infrequently consisted in the rejection of any connection between physicalism and the relevant counterfactuals.⁵⁰

Both difficulties mentioned in this section point in the same direction. The evaluation of an ordinary experiment is facilitated by reliance on powerful meta-justificatory principles concerning the reliability of observation and logic. In thought experiments, these principles don't apply. It appears that this substantially weakens the epistemic value of thought experiments.

Received: June 2005

Revised: April 2006

50 This is for instance the gist of John Perry's reply when he argues that physicalism isn't committed to the claim that Mary learns nothing new when she leaves her black-and-white room. See J. Perry, *Knowledge, Possibility, and Consciousness* (Cambridge, Massachusetts: MIT Press 2001).