THREE Thought Experiment

1. Introduction

A thought experiment is an experiment carried out in our imagination. It is a device used both in science and philosophy. In a thought experiment, we imagine a certain situation, we follow through some of the consequences of that situation, and then we draw a general conclusion — typically, a certain theoretical claim. A thought experiment is in some ways similar to an experiment in a physical laboratory. As a convenient label, call the latter kind of experiment a "concrete experiment." A difference between these two kinds of experiment is that a thought experiment concerns an imagined example whereas a concrete experiment concerns an actual example. But this contrast goes deeper. Concrete experiments are relatively familiar to us. (It is only relatively recently, though, that philosophy of science has recognized some of the complexities involved in experimental design and practice: see Gooding 1990.) But thought experiments seem more puzzling. The following three questions will be the focus of this chapter:

(Q1) A thought experiment is about an imagined situation. How can thinking about an imagined situation give new knowledge about the actual world?
(Q2) Are thought experiments special cases of a more familiar kind of phenomena, or are they *sui generis*?
(Q3) Are there particular difficulties facing philosophical thought experiments?

A thought experiment involves imagining a situation, not perceiving it. (Q1) asks how imagining a situation can tell us about what that situation would be like if it were actual, and what its theoretical consequences

would be. Furthermore, what kind of thing is a thought experiment? (Q2) asks whether we can understand thought experiments in terms of some more familiar, or better understood, ways. Lastly, thought experiments are of particular importance in philosophy because of many philosophers' epistemic ambitions. Many philosophers take themselves to have **modal knowledge** — knowledge about what is merely possible or about what is necessary. Some of them take thought experiments to be a fundamental source of modal knowledge. Some thought experiments seek to show that something imaginable is possible. Others seek to show that something that is apparently imaginable is impossible. Other philosophers, however, believe that there are special problems facing thought experiments in philosophy. Their criticisms chiefly concern how far removed the situations imagined in philosophical thought experiments are from actual situations, and the lack of background detail in these thought experiments. (Q3) asks whether any of these criticisms are good ones.

Before we embark on these large issues, it is well to stock up on lots of examples of thought experiments. That is the task of the next section.

2. Examples of Thought Experiments

IS THE UNIVERSE FINITE?

Imagine that the universe is finite and has a boundary. Now imagine throwing a spear at this boundary. Either the spear stops at the boundary or it passes through it. Lucretius argued that if the spear stops at the boundary, there must be something on the other side of the boundary that is stopping the spear, and so the so-called boundary is not a genuine boundary. Lucretius also inferred that if the spear passes through the so-called boundary, then again there is no genuine boundary. Either way space has no boundary. The thought experiment's conclusion is that space is infinite (Lucretius *The Poem on Nature: De Rerum Natura* p. 40).

NEWTON'S BUCKET

Imagine that a bucket filled with water begins to rotate. Initially, the surface of the water remains flat, but, as the water acquires the motion of the bucket, its surface becomes concave. The concave shape shows that the water is rotating. But the water is not rotating relative to the bucket because they have the same rotating motion. Newton infers that the water is rotating relative to absolute space (Newton 1686, book 1, *Scholium*).

DO HEAVIER BODIES FALL FASTER THAN LIGHTER ONES?

Galileo took Aristotelian mechanics to claim that heavier bodies fall faster than lighter ones. (There is some controversy whether this was

Aristotle's view, but we will let this pass for the sake of argument.) Galileo offered the following thought experiment as an objection. Imagine that a body *H* is heavier than a body *L*. Imagine further that *H* is connected to *L* by a cord. According to Aristotle, heavier objects fall faster than lighter ones. So will the composite body *H+L* fall faster than *H* or will it fall slower than *H*? On the one hand, given Aristotle's claim that heavier bodies fall faster than lighter ones, and since *H+L* is heavier than *H*, *H+L* will fall faster than *H*. On the other hand, since *L* is lighter than *H*, and again given Aristotle's claim that heavier bodies fall faster than lighter ones, *L* will retard the fall of *H* when they are joined together, and so *H+L* will fall slower than *H*. Conjoining these results, Aristotle's claim that heavier bodies fall faster than lighter bodies involves a contradiction (Galileo 1638, 66–68).

TRAVELLING AT LIGHT SPEED

Imagine what you would see if you travelled at the speed of light. According to Maxwell's theory of electrodynamics, you would observe the light beam as an electromagnetic field at rest. But since there is no such thing, according to Maxwell's theory, the thought experiment disconfirms the theory (Einstein 1949, 53).

THE MISSING SHADE OF BLUE

Imagine that you have not seen a certain shade of blue but that you have seen shades of blue slightly lighter or slightly darker than it. Could you imagine that "missing" shade of blue? Hume thought that you could. He concluded that not every simple idea we have needs to be a "copy" of a corresponding sense experience (Hume 1739-40, bk. I, pt. I, sec. I).

THE BRAIN IN A VAT

Imagine that your brain is placed in a vat wired up to a computer. The computer sends certain electronic signals directly to your brain thereby inducing experiences which are introspectively exactly like the experiences you would have if you were living a normal life and perceiving an external world. This imagined situation prompts the intuition that you cannot tell whether or not you are a brain in a vat. The conclusion is that you do not know whether you perceive an external world.

SWAPPED MEMORIES

Imagine that a cobbler lost all memories of his former life but apparently acquired all the memories that a prince has. John Locke inferred that the cobbler would be the same person as the prince, although they are different men (i.e., different human beings). Locke concludes that there

is a distinction between *being the same human being* and *being the same person*; these need not coincide (Locke 1694, bk. II, ch. 27, sec. 15).

THE CHINESE ROOM

Imagine that you are in a room with an input slot and an output slot. Through the input slot come messages in (say) Chinese. You have a manual that tells you what message in Chinese to post in reply through the output slot. The manual does not tell you what any of the Chinese messages mean in English; using the manual requires only that you match up the shapes of input Chinese characters with what's in the manual, and copy out characters just on the basis of their shape. John Searle has the intuition that you do not understand Chinese, even though you (unwittingly) give appropriate answers in Chinese to questions in Chinese. He concludes that understanding a language is not a matter of appropriate symbol manipulation, and, more generally, that the mind is not merely a symbol manipulating device analogous to a computer (Searle 1980).

MARY THE COLOUR SCIENTIST

Imagine that Mary is a colour scientist who has spent her whole life in a room in which everything is black and white and shades of grey. The physical sciences of her day are so advanced that their textbooks state all the physical facts involved when someone sees a red object. By reading these textbooks Mary learns all the physical facts involved when someone sees a red object. So Mary comes to know all the physical facts about what happens in people's nervous systems when they see red objects. According to physicalism, all the facts about what happens when someone sees a red object are physical facts. There are no non-physical facts involved. But when Mary leaves her room and sees a ripe tomato for the first time she will learn something that she did not know before. She will learn what a red object looks like. So Mary will learn a new fact about what is involved when someone sees a red object. But previously Mary knew all the physical facts involved when someone sees a red object. Therefore, this new fact that she learns is not a physical fact. It is a non-physical fact. So physicalism is false (Jackson 1982).

Lastly, here are briefer versions of some other philosophical thought experiments.

INVERTED SPECTRA

Imagine that I see as red everything that you see as green, and vice versa. Our behaviour and behavioural dispositions would be the same. So seeing things as red (or as green) cannot simply be a matter of behaving or

being disposed to behave in certain ways. This thought experiment for "inverted spectra" is in Locke (1694, bk. II, ch. xxxii, sec. 15), although the anti-behaviourist conclusion was drawn only after the rise of behaviourism in the 1930s.

ZOMBIES

Imagine an atom-for-atom replica of you that lacked any conscious experiences. This would be a physical replica of you that physically resembled you through and through and that behaved like you (a *Doppelgänger*), although it had no mental life. Such a physical replica would be a "zombie." Physicalism, however, says that you are nothing but a physical object. Since you differ psychologically from your zombie twin, the thought experiment concludes that physicalism is false (Chalmers 1996).

COULD THERE BE MORE THAN ONE SPATIAL WORLD?

You can travel from Toronto to Berlin. NASA can send a rocket from Earth to Neptune. More generally, it seems that there is a spatial route linking any two spatial regions. But is this a necessary truth? Imagine that every time that you fell asleep in your humdrum urban life, you then seem to wake in a sunny paradise with a different body from the one you have in your humdrum life. The people and sights around you are also quite unlike the ones in your humdrum life. And, whenever you seem to fall asleep at the end of a glorious day in this paradise, you awake again in your humdrum life. You set out on expeditions in your humdrum world but you never find the paradise, and vice versa. Anthony Quinton suggested that in this situation you would be experiencing life in two spatially unrelated worlds. He concluded that it is not a necessary truth that any two spatial regions are spatially connected (Quinton 1962).

MOLYNEUX'S CUBE

Imagine that someone blind from birth felt cubes and globes of about the same size. Imagine that that person later had his or her sight restored. Could that person tell just by looking which objects are cubes and which are globes? Molyneux, who devised the thought experiment, and Locke both thought that the person could not tell which was which. Locke concluded that our perceptions are altered by unconscious automatic inferences and that these inferences are due to our past experiences: "the ideas we receive by sensation are often in grown people altered by the judgment, without our taking notice of it" (Locke 1694, bk. II, ch. ix, sec. 8).

TWIN EARTH

Imagine a sample of liquid that had all manifest qualities of water, but which was not H_2O. (It might help to imagine this liquid to be found on another planet, Twin Earth.) Would that liquid be water? Putnam intuited that it would not. Consequently, since the word "water" would not apply to that liquid, the meaning of "water" is not determined by its manifest qualities alone. The microstructure of water, its being H_2O, determines what the word "water" correctly applies to. Since people may be ignorant of what the microstructure of water is, Putnam concluded that "'meanings' just ain't in the head" (Putnam 1975, 227).

THE VIOLINIST

Imagine that you found yourself wired up to a complete stranger whose life depended on the huge inconvenience of your remaining wired up to them for nine months. Would you have the right to sever the wiring? The situation you imagine yourself in is relevantly similar to a woman who finds herself pregnant. Judith Jarvis Thomson concluded that if you have a right to sever the wiring, by parity of argument the woman has a right to an abortion (Thomson 1971).

COULD EVERYTHING DOUBLE IN SIZE OVERNIGHT?

Imagine everything in the universe doubling in size at midnight. If you cannot imagine this, some philosophers think that this is evidence to show that objects are not located in absolute space like raisins in a pudding. Objects are simply in spatial relations to each other: there is no absolute space. (This is a variant of Leibniz's thought experiment for the relational nature of space: Leibniz 1715–16, 26.)

DUPLICATE PEOPLE

Imagine that two exact physical and psychological duplicates of Captain Kirk stepped out of the transporter room. Would those two people each be Captain Kirk? Parfit argued that they cannot each be Kirk since there is only one Captain Kirk, and it would be arbitrary to identify one of the duplicates with him. Parfit concluded that personal identity is a less important issue than has often been thought, and that what is more important is psychological continuity with past selves (Parfit 1984, 119–20, 282–87).

The variety of the above examples suggests to some people that thought experiments do not have a single role (Jackson 2009, 100–01). Some thought experiments clarify a theory. Other thought experiments clarify a consequence of a theory. Still others reveal otherwise unobvious

connections. And yet others provide test cases for philosophical analyses or scientific theories. Given that thought experiments may have any of these roles, the task of devising a theory of thought experiments becomes that much harder. The next section will review some attempts.

3. Theories of Thought Experiments

What kind of thing is a thought experiment? How does it work? In particular, how can we get new knowledge about the actual world just by imagining a situation? In this section we will outline five theories of thought experiments that seek to answer these questions.

(1) *Thought experiments as triggers* (Kuhn 1964). Anomalies, or results that conflict with prediction, turn up during scientific research. Often scientists ignore them by writing them off as experimental error or by hoping to tackle them at a later stage. The function of a thought experiment is to trigger scientists' memories of anomalies and thereby to retrieve knowledge of them. Imagining the situation presented by a thought experiment prompts scientists to remember the situation and what its consequences were. By drawing on their memories of the actual situation, scientists can reliably say what would happen if the imagined situation were actual. Kuhn claims that a thought experiment can precipitate a crisis in a scientific theory and so initiate a change of theory. By retrieving knowledge of an anomaly, scientists can spot some of the weaknesses of their current theory and then seek to change it.

Kuhn admits that his theory does not apply to all scientific thought experiments. Thought experiments describing physically impossible situations are exceptions (e.g., Einstein's moving at light speed). More generally, thought experiments describing unobserved types of situation do not fit Kuhn's theory. For example, Poincaré devised a thought experiment (his "Flat Land" thought experiment) which involved two-dimensional people living in a two-dimensional environment (Poincaré 1952, 37–38). Gendler (1998, 2000, 2004) develops Kuhn's idea that thought experiments can make us think about our theories in new ways, but she does so without relying on Kuhn's remembering account of thought experiments.

(2) *Thought experiments as a priori knowledge of a Platonic realm* (Brown 1991a, b, 2004a, b, 2007a, b). Some thought experiments are intellectual insights into a realm of properties that are not in space or time. These insights are like perceptions: they provide non-inferential *a priori* access to these properties and the lawlike relations between them.

This theory leaves the epistemology of thought experiments mysterious: how can minds get to know about things not in space and time? Brown makes a "companions in guilt" reply: it is equally mysterious how minds can get to know about things in space and time. Although it is relatively unmysterious how the external world acts on our nerve endings, it remains mysterious how changes in neurons produce changes in our beliefs (Brown 1991b, 65).

Even if this reply succeeds, the charge of mystery-mongering still has force. It is desirable to minimize the number of mysteries we admit. So a rival theory that does not harbour this mystery is, in that respect, a better theory.[1]

(3) *Thought experiments as arguments* (Norton 1991, 1996, 2004a, 2004b). Thought experiments are (inductive or deductive) arguments. The premises may be more or less explicit; the conclusion is the lesson that the thought experiment draws. The case for this theory is that various thought experiments can be represented as premise sets linked to conclusions by recognized forms of inference. Of course, the fact that thought experiments can be represented as arguments does not entail that thought experiments are arguments. But the fact that it is so illuminating and fruitful to represent thought experiments as arguments needs explaining, and the theory that they are identical provides a good explanation. For example, thought experiments can provide new knowledge because arguments can take us from familiar premises to surprising conclusions. The theory also makes thought experiments unmysterious. It is familiar and unmysterious that an argument can provide new and reliable information. An argument provides new information if its conclusion makes a claim that we do not already accept. The information provided is reliable if we have good reason to accept the argument's premise set. Now if a thought experiment is an argument, a thought experiment can provide new and reliable information in just the same way as an argument does.

Brown allows that some thought experiments are arguments, but denies that all are (Brown 1991b, 47). In particular, he denies that thought experiments such as Newton's bucket thought experiment are arguments. These are cases where the thought experiment establishes that there are certain phenomena (e.g., the states of the water before the bucket rotates and while rotating relative to the bucket), and we conjecture an explanation for the phenomena (here, the existence of absolute space) (Brown 1991b, 40–41).

1 For other criticisms of Brown's reply, see Norton (1993, 35–36), Sorensen (1992b, 1106–07), Cooper (2005, 333), and Häggqvist (2007).

But there is a straightforward way in which this kind of thought experiment can be construed as an argument: it can be construed as an inference to the best explanation. Newton thinks that certain phenomena need explaining; namely, the different states of the water in the bucket over time. He then selects what he takes to be the best potential explanation of this phenomena; namely, that in just one of these states the water and the bucket are rotating with respect to absolute space. Given the principle that it is warranted to believe that the best potential explanation of certain phenomena is the correct explanation, the thought experiment concludes that there is reason to believe that Newton's explanation is correct. (We discuss inference to the best explanation further in chapter 5, §5.)

Newton's thought experiment then has the following schematic argumentative form:

(1) Certain phenomena *p* need explaining.
(2) The best potential explanation of *p* is Newton's theory of absolute space.
(3) So probably Newton's theory of absolute space is true.

Brown writes that "absolute space is not the conclusion of an argument, it is the explanation for a phenomenon that Newton, in effect, postulates" (Brown 1991b, 48). But Brown makes a false opposition. The proposition *that absolute space exists* is both the conclusion of an argument — an argument that states an inference to the best explanation — and Newton's explanation of the behaviour of the water in the bucket.

Brown thinks that certain other thought experiments are not arguments. These are cases where the thought experiment starts with certain data and ends with a theory (Brown 1991b, 41–43). As an example Brown cites Galileo's thought experiment that all freely falling bodies fall at the same rate (Brown 1991b, 41). This is puzzling because earlier in his book Brown classifies this example as an argument; "it is a picturesque *reductio ad absurdum*" of Aristotle's theory of motion (Brown 1991b, 34). The thought experiment has the following schematic argumentative form:

(1) Heavier objects fall faster than lighter. (Assumption)
(2) Imagine two bodies, A heavier than B.
(3) Then A falls faster than B.
(4) So if A and B are united, then B will retard the fall of A.
(5) So the unit A + B falls slower than A.
(6) But the unit A + B weighs more than A.
(7) So the unit A + B falls faster than A. (Contradiction with 5.)

Since Aristotle's theory that heavier objects fall faster than lighter ones leads to a contradiction, similar reasoning shows that the theory that lighter objects fall faster than heavier ones also leads to a contradiction. But then it follows that freely falling bodies, whatever their weight, fall at the same rate. So Galileo's law of free fall is a corollary of his *reductio* argument against Aristotle. (Gendler [1998] claims that one of Galileo's thought experiments cannot be understood as an argument. For a reply, see Norton [2004b, §4.3].)

Another objection to the theory that thought experiments are arguments is that, since there are cases in which different parties reconstruct the same thought experiment as different arguments, such a thought experiment cannot be an argument (Bishop 1999 and Häggqvist 2009, 61). Notice, however, that there are also cases in which different parties reconstruct the same argument in Kant (his Transcendental Deduction) or in Wittgenstein (his anti-private language argument) as different arguments. That would not be a good objection to the claim that Kant and Wittgenstein offer arguments. In such a case, if none of the parties can be faulted on grounds of scholarship, it would be a misnomer to talk of *the* argument in the text. The author of the text was not sufficiently clear as to what his argument was, or perhaps different parties are sufficiently ingenious to read in lines of argument that did not occur to the author. A similar line can be taken with respect to thought experiments. In a case of the above kind, there is no unique thought experiment. There is an original statement that is suggestive of more than one thought experiment, each of which can be reconstructed as an argument.[2]

What we have seen so far is something of the versatility of the theory that a thought experiment is an argument. Although there may seem to be various kinds of thought experiments that are not arguments, this overlooks what forms an argument can take. As we have seen, an argument can be to the conclusion that a certain potential explanation is the best explanation of a phenomenon mentioned in the premise set. Given this, a thought experiment such as Newton's bucket can be readily accommodated as an argument.

Other objections to the theory are phenomenological. It has been claimed that running a thought experiment does not seem like giving an argument. "Thought experiments are often fun and easy, arguments are usually not" (Cooper 2005, 332). But "usually" is the giveaway. Why assume arguments are usually dull and difficult? (Bertrand Russell's were not.) And, that aside, why assume that thought experiments are like typical arguments? It is further objected that "when we perform a thought

2 Norton (2004b, 63–64) offers a similar reply.

experiment we imagine the situation unfolding in our mind's eye. We don't consider premises, modes of inference, and conclusions" (Cooper 2005, 332). Yet we are free to imagine a situation developing in any way we please. So why do we take it that an imagined situation would unfold (i.e., develop) in one way rather than another? The current theory offers an answer: given (a description of) an initial situation, and selected principles of development (certain inference rules), a certain further situation develops (a certain conclusion is inferred). Lastly, it is objected that a thought experiment such as Hume's one about the missing shade of blue "requires us to imagine what it is like to see blue, something that cannot be reduced to propositional form" (Cooper 2005, 332). Taken as an argument, Hume's thought experiment runs: "We can imagine a shade of colour if we perceive its neighbours in the colour spectrum. So we can imagine a colour shade if we perceive its neighbours, even if we have not perceived the shade being imagined." Let's grant that imagining what it is like to see the shade is not propositional. The argument does not assume otherwise. Instead, something propositional (the argument's premise) is used to represent something non-propositional (imagining what it is like to see the shade).

(4) *Thought experiments as genuine experiments* (Sorensen 1992a, Gooding 1990). The similarities between concrete experiments and thought experiments outweigh the dissimilarities. Both kinds of experiment can disconfirm theories, can identify interesting phenomena, and much else.

Nevertheless, emphasizing these similarities does not answer the principal question about thought experiments: how can imagining a situation tell us about what happens in actual situations? Since the theory at issue does not address that question, it is deficient. The theory can be supplemented with an epistemic account. Sorensen, for instance, claims that thought experiments are both experiments and arguments: they involve "a set of individually plausible yet inconsistent propositions" (Sorensen 1992a, 6). Given this combination of theories, it is unclear what work the "thought experiments are genuine experiments" theory does. Gooding talks of the "construction of experimental narratives that enable virtual or vicarious witnessing" (Gooding 1990, 204–05). Gooding does not develop those remarks and, as they stand, they can be interpreted in terms of any of the other theories of thought experiments. For example, on Kuhn's theory of thought experiments as triggers for memory, presenting a thought experiment can be understood as developing a story about an imaginary experiment, where hearing this story ("witnessing

it") triggers a memory of that experiment actually being performed and producing a certain anomalous outcome.

(5) *Thought experiments as models of possible worlds* (Nernessian 1991, 1993, Miščević 1992, and Cooper 2005). Thought experiments pose "what if" questions. What would happen if bodies obeyed Aristotelian mechanics? What would happen if Mary saw something red for the first time? To answer such questions we predict how these objects would behave in the imagined circumstances. In some cases (such as the Aristotelian case), we know which laws would govern objects in the imagined circumstances, and we can thereby predict the objects' behaviour. In other cases, we can use our implicit understanding of laws that we cannot fully state. In both cases, what knowledge we have of these laws enables us to develop a model — a representation of various possible situations.

The theory that a thought experiment is a model of a possible world does not help answer any of the epistemological questions about thought experiments that have already been raised. Our key original question was: how does imagining something give new knowledge about the world? The present theory faces a variation of this question: how does devising a model give us new knowledge of the world? The fact that a given model is consistent (or impossible) tells us that it is consistent (or impossible) for the world to be that way only on the assumption that the model is an accurate model of the world. We cannot always tell from our armchairs, however, when that assumption is correct. For example, we cannot tell from our armchairs what laws of nature hold. Perhaps what the theory leads to is the view that a thought experiment articulates a counterfactual claim: if such and such a model were an accurate model of the world, then so and so would be the case.[3]

4. Scepticism about Philosophical Thought Experiments

Scepticism about thought experiments in philosophy stems from a number of quarters. We will review these criticisms in a series of sub-sections.

WHAT KIND OF REASONS DO THOUGHT EXPERIMENTS PROVIDE?

If both a concrete experiment and a thought experiment can provide epistemic reason to believe a scientific theory *T*, what kind of epistemic reason is this? A concrete experiment can show that *T* makes a correct prediction by testing of one of *T*'s predictions. By contrast, a thought

3 For a similar view, see Williamson (2007, ch. 6), but see Ichikawa (2009) for criticisms.

experiment does *not* test the predictions of a theory. "[T]he function of thought experiments in science is to draw out the physical implications of our theories and to test their nonempirical virtues" (Bokulich 2001, 303). The theoretical (or "nonempirical") virtues of a theory include its explanatory power, its simplicity, its consistency, and its fruitfulness (its ability to suggest novel hypotheses). There is an issue about what the significance is of the fact that a given theory has a certain theoretical virtue. In particular, is it a reason to believe that theory? (We will take this issue up in the case of simplicity in chapter 4, §6.) Perhaps thought experiments provide no reason to believe (or disbelieve) a theory but play only a popularizing or heuristic role in presenting the theory, its commitments, and how it might be tested by actual experiments.

DO THOUGHT EXPERIMENTS GENERATE CONTRADICTORY CONCLUSIONS?

Jeanne Peijeneburg and David Atkinson claim that disagreement about the conclusion of a given thought experiment indicates that the thought experiment is a poor one (Peijeneburg and Atkinson 2003, 308–10). They think that this exposes many philosophical thought experiments as poor ones: "thought experiments in contemporary analytic philosophy often generate contradictory conclusions" (Peijeneburg and Atkinson 2003, 308). For example, one philosopher might conclude from the thought experiment about Mary the colour scientist that physicalism is false. Yet another philosopher might instead claim that it is psychologically impossible for someone to know all the physical facts about the working of the brain, and conclude that no lesson about physicalism can be drawn from the thought experiment. And still another philosopher might claim that were Mary to know all the physical facts about colour, then she *would* know what red things look like.

Peijeneburg and Atkinson admit, however, that this point also holds for certain scientific thought experiments, such as Newton's bucket thought experiment (Peijeneburg and Atkinson 2003, 306). Their response is to say that a scientific theory can provide reason to believe a particular conclusion of a scientific thought experiment (Peijeneburg and Atkinson 2003, 315). Their idea seems to be that if scientific theory T_1 is better than a rival T_2, then we should believe the conclusion that T_1 draws from a thought experiment rather than the conclusion that T_2 draws (if those conclusions differ). But then it seems that a similar principle can be used to select between conflicting conclusions drawn from philosophical thought experiments. Peijeneburg and Atkinson reject this, remarking that "[in] philosophy, however, the turn to theories is of little help. How should we decide between, say, the theories of Searle and Dennett on

understanding, meaning and consciousness?" (Peijeneburg and Atkinson 2003, 315). Their argument can be reconstructed as follows:

(1) There is no reason to believe one of those philosophical theories rather than the other except for the degree of support it gets from thought experiments.
(2) The only reason to believe one conclusion of philosophical thought experiment rather than a rival conclusion would be because of a reason to believe one of these philosophical theories rather than another.
(3) So there is no reason to believe one conclusion of a philosophical thought experiment rather than a rival conclusion.

It follows that we cannot work out which thought experiments are good ones (and so which philosophical theories are good ones) on pain of circularity.

The above argument is valid. The trouble with the above argument is that premise (1) is very contentious and Peijeneburg and Atkinson provide no justification for it. And if the premise itself states a philosophical theory — the philosophical theory that the only source of justification for such a theory is a thought experiment — then they cannot provide justification for it, on pain of contradiction.

Furthermore, what Duhem taught us about concrete experiments applies with equal force to thought experiments (Duhem 1914, 188-90, 204).[4] When a theory faces putative disconfirmation from a thought experiment or from a concrete experiment, it is always possible to modify the theory to avoid disconfirmation. For example, instead of taking the theory to be false, perhaps we should take some of the background assumptions used in testing the theory to be false. Or perhaps we wrongly assumed that certain potentially interfering factors were absent. Or again perhaps we wrongly assumed that certain factors were innocuous when in fact they interfered. If we made any of these revisions, we need not take the theory to be false. The issue then is whether the costs of the revision exceed the benefits: is the revision purely *ad hoc*, does it make the theory less simple, or does the theory make a more than compensating gain in explanatory power? Second, a thought experiment is always open to interpretation, and two rival theories may offer different interpretations of the same thought experiment. In this case, what the

4 See Bokulich (2001, 288–89) for the extension of Duhem's point to thought experiments.

best interpretation of the thought experiment is will be underdetermined. This issue also arises in the next sub-section.

ARE THOUGHT EXPERIMENTS QUESTION-BEGGING?

Peijeneburg and Atkinson also claim that another indicator that any given thought experiment is a poor one is if it assumes the very intuition that it is supposed to elicit (Peijeneburg and Atkinson 2003, 311). We will see an alleged illustration shortly. Their claim seems to be a special case of the general point that any argument that begs the question is defective. "The conclusions drawn from thought experiments beg the question: they hinge on intuitions of which the truth or falsity was supposed to be demonstrated by those very thought experiments" (Peijeneburg and Atkinson 2003, 310).[5] An argument begs the question when it contains at least one premise that would not be accepted by the target audience because they do not yet accept the conclusion of the argument. Or, more simply, anyone would have reason to accept all of its premises only if that person has independent reason to accept its conclusion (Walton 1989, 52 and Govier 1992, 85). Begging the question is a defect in any piece of reasoning. But is there any reason to think that philosophical thought experiments especially suffer from this defect?

Consider Quinton's thought experiment (see §2). Quinton imagines your having various experiences while awake in your humdrum life and various more exotic ones while asleep in your humdrum life. He interprets this as a situation in which you have experiences of two spatially unrelated worlds, the humdrum world and the exotic paradise. He concludes that it is not a necessary truth that every space is spatially related to every other space. But Quinton's interpretation assumes that it is possible for two worlds to be spatially unrelated — and that is the very conclusion to be established.

But just because Quinton's thought experiment has this failing, it cannot be supposed that every philosophical thought experiment does. A thought experiment does not beg the question just because it expresses an intuition in its conclusion. The premises of a thought experiment may give reason to accept the conclusion, and so give reason to accept the intuition. That is a perfectly reasonable way to proceed. Without building in the intuition as one of the premises of the argument, the argument shows that the intuition follows from a set of premises that there is independent reason to believe.

Peijeneburg and Atkinson also say that the conclusions of thought experiments beg the question because "they are embodiments of those

5 See also Ward (1995).

intuitions for the sake of which the entire thought experiment [*sic*] was conceived" (Peijeneburg and Atkinson 2003, 317). But this confuses the motivation for giving an argument with the question of whether the argument is question-begging. An argument may be given to show that *p* is true. It does not follow that the argument begs the question as to whether *p* is true. For example, suppose I want to persuade you that Pele is rich. I might argue as follows: all world class footballers are rich; Pele is a world class footballer; so Pele is rich. I gave that argument in order to persuade you that Pele is rich. But the argument did not beg the question as to whether he is rich.

Lastly, if Peijeneburg and Atkinson's claim that philosophical thought experiments beg the question were correct, their claim would apparently generalize so that all thought experiments beg the question. So scientific thought experiments would have the same defect. The authors think that this consequence can be avoided because concrete experiments can be performed in science to test the thought experiment's conclusion (Peijeneburg and Atkinson 2003, 317). But, first, even if the conclusion of a thought experiment can be tested, it is hard to see how that makes the thought experiment a good one. If argument *A* begs the question, the argument is not made a good one if a non-question begging argument *B* can also be provided for the same conclusion. Argument *A* remains a poor argument. Likewise, if thought experiment *C* begs the question, it is not made a good one if a concrete experiment can be conducted to reach the same result as the thought experiment predicted.

Second, the conclusions of some good scientific thought experiments cannot be tested in this way because they concern situations that are physically impossible. Einstein's thought experiment of what someone travelling at the speed of light would observe is one such example.

Third, the conclusions of some philosophical thought experiments can be tested by concrete experiments. Searle's Chinese room and Molyneux's example could all be carried out in real world experiments. Peijeneburg and Atkinson reply that such experiments "would not resolve the philosophical conundrum" (2003, 317). But now it seems that more is being required of a good thought experiment than that its conclusion is testable in a concrete experiment. Here it is being further required that the experiment resolves the philosophical dispute. The requirement seems unreasonable, however, given what was said above about Duhem's thesis and its implication for thought experiments. No concrete or thought experiment can be guaranteed to resolve a dispute between theories.[6]

6 See also the exchange between Cohnitz (2006) and Peijeneburg and Atkinson (2006).

We now have some sense of sceptical arguments against thought experiments in philosophy. There are further such arguments to address. As we have found in our two earlier chapters, the most profitable way to pursue a methodological issue is by examining how that methodology applies to a particular case study. To this end, we will consider a case study in §5. This will concern scepticism about thought experiments about personal identity. §6 will discuss whether philosophical thought experiments can be tested by empirical means. A recent movement, **Experimental Philosophy**, champions the view that it is very useful to do such empirical testing. Some of its findings have been especially interesting because they have been markedly negative.

5. Case Study: Thought Experiments about Personal Identity

One theory of personal identity identifies a person with his or her own body. The following brain transplant thought experiment is designed to challenge that theory. It is possible that a person's entire brain is transplanted into a new skull (or into a suitably prepared laboratory vat), and kept alive so that its brain functions continue as before. The brain would then remain conscious: it would produce experiences, thoughts, beliefs, apparent memories, and so forth. Indeed, there would seem to be psychological continuity between the brain's mental states before, and after, the transplant. This may prompt the intuition that where the brain goes, the person goes. Now let's assume that either the person is identical with the original body minus the brain, or the person is identical with the brain. Suppose that, perhaps thanks to the thought experiment, you have the intuition that the person is located wherever the brain is located. So, following the transplant, the person is not located where the original body is. It is a plausible principle governing identity that x is identical to y only if, at any time, the location of x is identical to the location of y (if x or y are located). It follows that the person is not identical to the original body. The brain transplant thought experiment is potentially a very powerful device. If successful, it would show that the theory that a person is identical to his or her body is false.

What should we make of such a thought experiment? Can it really undermine its intended target? One of the principal claims of Kathleen Wilkes's book *Real People* is that the practice of discussing personal identity in terms of what she calls "theoretically impossible" speculative cases is misguided (Wilkes 1988, ch. 1). Mark Johnston and Peter van Inwagen have similar views (Johnston 1987; van Inwagen 1997, 307–08).

Wilkes and Johnston each run the following line of argument. There is no background of theory against which such speculations can be evaluated. Consequently, either we run a thought experiment against the background of what we already believe about the world, or we run it against a quite different background of beliefs. On the first limb of the dilemma, we are running the thought experiment against a background that rules it out as "theoretically impossible," i.e., as conflicting with what we believe the laws of nature to be. On the second limb of the dilemma, we have an idle fantasy on a par with imagining that there are carnivorous rabbits on Mars, or that people can pass through mirrors.

Wilkes's positive proposal is that discussions of personal identity should consider only human beings and what actually happens to them. What occurs to some human beings is sufficiently puzzling and thought provoking that wildly speculative cases can be set aside without loss.

Let's consider the first limb of the above dilemma. James Robert Brown replies that:

> Too often thought experiments are used to find the laws of nature themselves; they are tools for unearthing the theoretically or nomologically possible. Stipulating the laws in advance and requiring thought experiments not to violate them would simply undermine their use as powerful tools for the investigation of nature. (Brown 1991b, 30)

Wilkes might reply that her claim is also consistent with Brown's view that thought experiments are epistemic tools for discovering laws of nature. Suppose that we make a series of thought experiments designed to find out the laws of nature. Suppose that our first thought experiment discovers that *L* is a law of nature. Wilkes's claim is that, given this discovery, no subsequent thought experiment should make a claim incompatible with *L*. The judgement that *L* is a law could not rationally be revised on the basis of a thought experiment.

Although this point shows that Brown's view is consistent with Wilkes's claim, it is itself open to objection. Some scientific thought experiments concern situations that are "theoretically impossible" in Wilkes's sense. Recall Einstein's thought experiment about what he would observe if he travelled at the speed of light. This thought experiment remains highly regarded despite the fact that it is theoretically impossible (i.e., incompatible with the laws of nature) for Einstein to travel at the speed of light — as Einstein himself recognized. Accordingly, Wilkes's claim flouts best scientific practice.

What seems right about Wilkes's claim is that, if we are given a description of a situation that we cannot make much sense of, anything further that we say about the situation will be guesswork. Consequently, anything further we might say will not count as evidence for, or against, any of our beliefs. This does not, however, license a blanket ban on philosophical thought experiments. To repeat: the ban applies only to descriptions of situations that we do not understand. And saying which descriptions these are should not be done in a casual matter. We have to think each description through on a case by case basis. If we fail to understand a description, we should at least have tried to understand it in the first place (Kitcher 1978, 105). None of the sample thought experiments given earlier seems to ask us to imagine the nonsensical. One point of some thought experiments might be to show that something that looks prima facie to be imaginable is actually impossible or incoherent.

Why is Wilkes sceptical about philosophical thought experiments? She has two arguments. Her first argument is as follows. If many philosophical arguments are judged against the world as we know it, they describe "theoretical" or "in principle" impossible situations. Therefore, they are no more than fantasy. Examples of "theoretical" impossible situations include gold not having atomic number 79, or water not being H_2O. Gold and water are examples of natural kinds: natural collections of natural things. Other natural kinds include tigers, spiders, and roses. According to Wilkes, thought experiments involving a natural kind are unsuccessful if they take instances of that natural kind to lack any of their essential properties. Gold is a natural kind, and its instances essentially have atomic number 79. Water is a natural kind and its instances are essentially composed of H_2O molecules. Hence thought experiments in which gold or water lack these, or any other, essential properties are unsuccessful. Wilkes further claims that human beings form a natural kind. Consequently, she claims that a thought experiment about human beings is unsuccessful if it takes a human being to lack any of its essential properties. The issue then is whether thought experiments such as the brain transplant thought experiment take a human being to lack any of its essential properties.

It might be thought that Wilkes's argument can be side-stepped. The response runs as follows. It seems possible that there are persons who are not human beings (as Wilkes [1988, 36] apparently concedes). If so, the philosophical thought experiments about persons that Wilkes objects to can be re-cast in terms of non-human beings who are persons. By re-casting the thought experiments in these terms, Wilkes's objection is circumvented (Madell 1991, 139). Thought experiments about persons who are not human beings are not thought experiments rooted in observation

and experience. On the face of it, they are thought experiments about purely imaginary beings. This, however, takes us to Wilkes's second argument against philosophical thought experiments.

Her second argument runs as follows. All thought experiments make background assumptions. In scientific thought experiments, these background assumptions are made explicit. Many philosophical thought experiments, by contrast, leave their relevant background assumptions unspecified and only implicit. In this respect these thought experiments resemble fairy tales. But while, for the sake of an entertaining story, we may waive *how* Alice can pass through a mirror, philosophical thought experiments need to be more detailed and rigorous. Such thought experiments need to explain how, for example, a person's brain could be successfully transplanted from one body into another in a way that preserves that brain's mental and other functions. Again, the thought experiments need to explain how a person could be teletransported from one place to another.[7]

Taking the second of these examples, Madell replies that Wilkes's argument at most shows that we do not know how someone could be teletransported, but it does not show that the thought experiment is incoherent (Madell 1991, 139). Wilkes might respond that Madell's reply misses the point. The charge was not that the teletransporter thought experiment was incoherent. The charge was that, unless we are told how the thought experiment is possible, there is no more warrant for thinking that the thought experiment is coherent than that there is warrant to think that the fiction of Alice stepping through a mirror is coherent. Unless this challenge is met, there is no more reason to think that the thought experiment is coherent than there is to think that any fairy tale that does not contain an overt contradiction is coherent.

On which side does the burden of argument lie? Is Madell obliged to offer more justification for accepting the teletransporter thought experiment as coherent? Or is Wilkes obliged to offer more justification for questioning its coherence? Madell thinks that the onus is on Wilkes to "justify rejecting thought experiments [that are not overtly contradictory]" (Madell 1991, 139).[8] Lycan puts the point more generally:

> For any modal claim that something is a necessary truth, I would say that the burden is on the claim's proponent. A theorist who maintains of something that *is* not obviously impossible that nonetheless that thing is impossible owes us an

7 For related worries, see Cooper (2005, 345).

8 See also Snowdon (1991, 115).

> argument. And since entailment claims are claims of necessity and impossibility the same goes for them….
>
> … The proponent of a necessity, impossibility, entailment or incompatibility claim is saying that *in no possible world whatever* does it occur that so-and-so. That is a universal quantification. Given the richness and incredible variety of the pluriverse [i.e., the plurality consisting in every possible world], such a statement cannot be accepted without argument save for the case of basic logical intuitions that virtually everyone accepts. (Lycan 2003, 109)

To talk about "the richness and incredible variety of the pluriverse," however, does not settle matters. The issue is just how rich the "pluriverse" is. Granted the pluriverse contains every possibility. It is still up for argument what is possible and what is not.

Disputes about where the burden of proof lies often end in stalemate. When thought experiments describe fantastical situations, Madell and Snowdon take the default view to be that it is "Business As Usual." For Johnston and Wilkes, however, it is a case of "All Bets Are Off." Imaginative philosophers are free to dream up cases at will, and to stipulate that they are cases of (say) teleportation. Johnston and Wilkes warn that that does not make them cases of metaphysical discovery. For these critics, devising fantastical thought experiments "simply fails to make any contact with reality, and it is hard to see why discussions of such cases should be of any interest to [metaphysics]" (Maudlin 2007, 188).

These critics might run the following argument (Stroud 1977, 50). Some mathematical claims have been neither proved nor disproved. It is an open question which of these claims states something necessarily true, and which of them states something necessarily false. Goldbach's conjecture is the claim that every even number greater than two can be expressed as the sum of two primes. Now it seems that there is no overt contradiction in imagining proving the conjecture. Perhaps you imagine yourself working long and hard, scribbling down equations until you conclude that Goldbach's conjecture is true. There is also no overt contradiction in imagining disproving Goldbach's conjecture. Yet these thought experiments cannot both be successful. They cannot each reveal a genuine possibility because it is possible to prove Goldbach's conjecture only if it is impossible to disprove the conjecture. So a thought experiment that is not overtly self-contradictory may still describe something impossible.

Some philosophers think that mathematical claims that are neither provable nor disprovable are neither true nor false. We need not quarrel with these philosophers here. Even if Goldbach's conjecture is neither provable nor disprovable, the above point still holds. A thought experiment that describes you proving the conjecture and a thought experiment that describes you disproving the conjecture cannot both describe genuine possibilities. Perhaps neither thought experiment describes a genuine possibility because the conjecture can neither be proved nor disproved.[9]

The above line of argument can be challenged (McGrew and McGrew 1998, pt., 3). First, it is questionable whether, by imagining the above situations, you genuinely imagine Goldbach's conjecture being true or genuinely imagine it as false. Strictly what is imagined in the first situation is that you have written down a string of equations and form the belief that you have proved Goldbach's conjecture. That falls short of proving the conjecture: it is consistent with what you have imagined that the conjecture is false. Similarly, imagining your believing — even your justifiably believing — that the conjecture is false falls short of imagining that the conjecture is false. Notice that there is no contradiction between imagining your believing the conjecture *and* imagining your disbelieving the conjecture.

Madell and Snowdon might further reply that, when there is no evidence of any kind that a given thought experiment is impossible, the default view should be that the thought experiment describes a genuine possibility. If this suggestion is to be made tenable though, it would need to be formulated more carefully. Since some thought experiments, such as Quinton's, are unsuccessful, and so thought experimenting is not a fail-safe epistemic method, there is some evidence against any given thought experiment being successful. There is also a question of what counts as evidence against a given thought experiment. Does imagining that every attempt to build a teleportation device fails count as evidence that teleportation is impossible? Would that thought experiment conflict with Parfit's thought experiment much as the above pair of thought experiments about Goldbach's conjecture were conflicting thought experiments? It would not be helpful to suggest that, unless there is sufficient evidence against a thought experiment, the default view should be to take the thought experiment to describe a genuine possibility. We need to be told how much evidential support the default provides, and so what degree of counter-evidence would be needed to defeat it.

9 For further discussion of this issue, see Yablo (2000), Gendler and Hawthorne (2002, introduction), and Rosen (2006).

In fact, it is doubtful whether there is any methodological default for or against possibility claims. To say that it is possible that a donkey talks is to say that it is consistent with the nature of a donkey that it talks. To have a reason to believe such a claim, or to have a reason to believe such a claim is false, depends crucially on what one knows, or has good reason to believe, about that nature. So, in any given case, the key question to ask is what those reasons are. This was the point that Wilkes made in her emphasis on what natural kinds a given thing belongs to. Judgements about what is possible or what is impossible have to be assessed on a case-by-case basis, and not by some burden-shifting rule.

What is meant by a background assumption to a thought experiment? Do philosophical thought experiments leave their background assumptions unclear? If so, does this tell against those thought experiments? Can the background assumptions be made clear? If they are made clear, do any other difficulties face philosophical thought experiments?

It may be that philosophical thought experiments often leave their background assumptions unclear in the sense that they do not explain how those assumptions might come about, or they fail to explain all of the important consequences of those assumptions. But then the same goes for the scientific thought experiments (Snowdon 1991, 120). Take Einstein's thought experiment of what someone would observe if they travelled at the speed of light. Einstein does not seek to explain how the observer could travel at that speed, what the means of propulsion would be, how the observer could survive the resulting increase in mass, and so forth. What we have here is a "companions in guilt" reply to Wilkes. Wilkes makes a certain objection against philosophical thought experiments because they fail to explain how their assumptions come about, or what all their important consequences are. It is then replied that at least some scientific thought experiments have the same features. So either these scientific thought experiments are as bad as the philosophical thought experiments are claimed to be — both kinds of thought experiment are "companions in guilt" — or there is nothing objectionable about thought experiments having the features in question — both kinds of thought experiment are "innocent." Since Wilkes wants not to impugn scientific thought experiments, it seems that her current objection to philosophical thought experiments fails.

To sum up this section, there is no default assumption that a thought experiment describes a genuine possibility unless shown otherwise. Thought experiments need to be assessed on a case-by-case basis.