WHEN ARE THOUGHT EXPERIMENTS POOR ONES?

JEANNE PEIJNENBURG and DAVID ATKINSON

SUMMARY. A characteristic of contemporary analytic philosophy is its ample use of thought experiments. We formulate two features that can lead one to suspect that a given thought experiment is a poor one. Although these features are especially in evidence within the philosophy of mind, they can, surprisingly enough, also be discerned in some celebrated scientific thought experiments. Yet in the latter case the consequences appear to be less disastrous. We conclude that the use of thought experiments is more successful in science than in philosophy.

Key words: EPR, Kant's antinomies, Newton's bucket, thought experiments

1. INTRODUCTION

One of the things that sets contemporary analytic philosophy apart from its older variant is its ample use of thought experiments. Whereas early analytic philosophers like Russell, Ayer or Carnap seldom rely on this kind of hypothetical reasoning, modern ones like Jackson, Searle and Putnam do not eschew the most bizarre accounts of zombies, swapped brains, exact *Doppelgänger*, and famous violinists who are plugged into another body. Many people have expressed feelings of unease when confronted by these outlandish stories (Quine, Dennett, Wilkes, Häggqvist, etc.). This unease becomes acuter when one compares the grotesque stories with thought experiments that are successful, such as that by which Galileo Galilei disproved the Aristotelian theory of falling bodies. There seems to be a world of difference between Galileo's exemplary argument on the one hand and such a far-fetched story as Searle's Chinese Room on the other.

A world of difference, but what exactly *is* that difference? *Why* do we have the feeling that Galileo's argument is decisive, whereas Searle's argument only takes us further from home? The question touches the criteria which distinguish good thought experiments from bad ones. Without professing to list those criteria in full, we claim that we can formulate two

Journal for General Philosophy of Science **34:** 305–322, 2003.

indications on the basis of which we might suspect a particular thought experiment of being a poor one. We shall discuss the first indication in Section 2 and the second in Section 3. Unlike most sceptics on philosophical thought experiments, we shall argue that both indications apply to some celebrated scientific thought experiments as well (Sections 4 and 5).

Since we are preoccupied with the difference between good and bad, we do not feel the need to state exactly what thought experiments are; after all one can distinguish good from bad theories, or thoughts, or experiments without being able to define what exactly theories, thoughts or experiments are. Not that there is a dearth of definitions. On the contrary, a lively debate on the nature of thought experiments can be discerned in the literature from 1990. Thought experiments have been defined as limiting cases of experiments (Sorensen 1992), as arguments (Norton 1991, 1996), as 'guided contemplations' (Gendler 1998), as vistas in a Platonic world (Brown 1991), as specific functions in an experiment (Borsboom *et al.*, 2002), or as not arguments at all (Bishop 1999). However, this disagreement about what thought experiments *are* contrasts with the unanimity about what thought experiments should *do*; and it is the latter, not the former, that counts if we want to distinguish good from bad thought experiments.

In the view of almost everyone, a thought experiment should give sudden and exhilarating insight. Very often, as many have observed, the insight so achieved amounts to seeing that an existing theory is false (Brown 1991, Sorensen 1992, Norton 1996). The setting of thought experiments that are constructed with the aim of refuting a theory – 'destructive' thought experiments, in the locution of Brown – can be symbolized as follows:

$$(T \,\&\, E) \rightarrow S, \neg\, S, E \vdash \neg\, T$$

where T is a theory, E symbolizes a thought experiment, and S is a particular situation of which everybody knows that it is not the case, i.e. $\neg$ S.[1] Searle's Chinese Room experiment may serve as an example: from strong Artificial Intelligence (theory T) together with the Chinese Room experiment (E), it follows that Searle, locked up in the room, understands Chinese (S). But everbody knows that Searle does not understand Chinese ($\neg$ S), therefore strong A.I. is false. Putnam's Twin Earth constitutes another example: from meaning internalism (T) together with the Twin Earth experiment (E), it follows that the term 'water' on Twin Earth denotes H_2O (S). However, it is given that 'water' on Twin Earth denotes XYZ, hence 'meanings ain't in the head'. A third example we find in the case of Frank Jackson's Mary. The supervenience thesis and the Mary experiment together imply that Mary does not learn anything new when she finally leaves her black-and-white room; but since it is clear that she does (she learns what it is like to see red), the supervenience thesis is incorrect.

Even Galileo's famous Pisa experiment can be seen in this setting: from the Aristotelian theory of falling bodies and the Pisa experiment it follows that a musket shot tied to a cannon ball falls faster *and* slower that a musket shot alone. Everybody knows that this cannot be so ($\neg S$), thus Aristotle is wrong. In all these cases, one tries to refute an existing theory with the help of a counterfactual situation and a particular (e.g., a particular person, standing on a particular tower in Pisa, drops a particular musket shot, etc.)[2]

A diligent user of thought experiments in analytic philosophy, Derek Parfit, has defended their frequent use by stating that thought experiments arouse in us 'strong beliefs' (Parfit 1984, 200). These strong beliefs correspond to $\neg S$ in the formula above: everybody knows that S is false, so $\neg S$ is a strong belief. It is notably through these strong beliefs that we gain what thought experiments are supposed to deliver: sudden and clear insight (for instance that a certain theory is wrong).

The strong beliefs may be false, and Parfit admits as much. However, Parfit neglects to caution us that often the strong beliefs also contradict each other. Yet it is the latter fact that poses a real threat to thought experiments. For a thought experiment can only be deemed successful if it induces the same – true or false – belief in the majority of people that are exposed to it. Nobody doubts that true beliefs are better than false ones, but much more significant than its truth or falsity here is the fact that the *same* belief is induced in almost everybody who is exposed to the experiment at hand. It is this 'collectivity' that does the trick, and a thought experiment that prompts diametrically opposed beliefs is not very successful, to say the least. (In this respect, thought experiments resemble logical puzzles such as the Monty Hall Dilemma. Here a quiz master confronts us with three doors, one of which hides a fiercely desired prize. We choose a door, say door 1, but instead of opening it the quiz master opens a door that does not hide the prize, say door 2. Question: is it advantageous to change to door 3? The reason why these puzzles are so disturbingly successful is because they arouse in almost all of us the same – false – belief. Were it the case that these puzzles triggered radically *different* rather than just false beliefs, so that in the end some people acquired a strong belief in X whereas others, equally strongly, believed *not-X*, the puzzles would be rather pointless. If, say, fifty percent of the audience in the Monty Hall show believed that switching to door 3 would increase the chance of winning the prize, whereas the other fifty percent believed that it does not make any difference whether one switches or not, the programme's ratings would not have been as high as they were. But in fact most people's first reaction is the same: they – incorrectly – conclude that changing does not make any difference.)

Nevertheless, as we will see in Section 2, contemporary analytic philosophy contains more than one example of a thought experiment that has given rise to just such conflicting reactions. In other words, a strong belief in $\neg S$ turns out to be all but self-evident: some people have, on the contrary, a strong belief in S.

2. FIRST INDICATION: CONCLUSIONS CONTRADICT ONE ANOTHER

James McAllister has argued that in science thought experiments did not really occur before the time of Galileo (McAllister 1996). Galileo realised that some of his experiments, when actually performed, would yield contradictory conclusions. Take again the Pisa-experiment, which, as scholars agree, Galileo never actually carried out. If you really start dropping cannon balls and musket shots from the leaning tower, you will notice that repetition of the experiment will yield different outcomes each time: sometimes the cannon ball arrives first, sometimes the musket shot arrives first, sometimes they arrive simultaneously. The causes of these differences are of course all sorts of unforeseen factors: the wind, the air friction, the uneven ground, and, last but not least, the difficulties you will encounter in releasing the objects together. In order to avoid such accidental factors, McAllister argues, Galileo performed the experiment only in thought. For only by restricting himself to a thought experiment could Galileo reveal the laws of falling bodies.

McAllister's argument gives an ironic twist to our point: whereas he claims that scientific thought experiments came into being with the aim of evading the contradictory conclusions with which ordinary experiments would often have left us, we claim that thought experiments in contemporary analytic philosophy often generate contradictory conclusions. Two examples may serve here.

Think of all the thought experiments about 'twin-you', your exact replica or *Doppelgänger*. By definition, you and twin-you are physically, 'molecule-for-molecule', the same. Question: are you and twin-you also mentally identical? Will twin-you, as Kim puts it, 'be as smart and witty as you, as prone to daydream, share your likes and dislikes in food and music, and behave just as you when angry? ... Will his twinges, itches, and tickles feel to him just the way yours feel to you?' (Kim 1996, 9–10). Yes, of course they will!, is the straightforward answer of Davidson, Hellman & Thompson, Dennett, Burge, Papineau, and all the defenders of the supervenience thesis in one of its different versions. No, of course they won't! is the heartfelt answer of philosophers like Thomas Nagel, Frank Jackson, or David Chalmers. According to Chalmers, twin-you is a zombie, a creature that is physically and perhaps even psychologically identical to you (in the sense that it can functionally perceive trees outside, recognize the taste of chocolate, or report the contents of its internal states), but that will never be your duplicate in the 'phenomenological' sense. For in zombies, none of the physical or psychological functioning will be accompanied by any real conscious experience: 'There will be no phenomenological feel. There is nothing it is like to be a zombie' (Chalmers 1996, 95).

Thus thought experiments about physical replicas trigger two mutually inconsistent intuitions. On the intuition of the defenders of the supervenience thesis, my physical clone is also my mental clone, since there is no mental difference without a physical difference. But on the intuition of Chalmers *et al.*, the very conceivability of non-conscious creatures that are like me in every physical and functional respect shows that there can be mental differences without physical differences. Which intuition is the right one? There is no way in which we can answer that question, for we do not know what *is*, and what is *not* implied by the idea that physical duplicates of us are walking around. In particular no-one knows whether his physical duplicate is numerically identical to him, and thus would indeed be he. The best thing we can do, so it seems, is to devise another thought experiment. But summoning a thought experiment in order to resolve a thought experiment is a doubly dubious undertaking.

Another example: the remarkable story of Mary-the-colour-scientist. Twenty-three year old Mary knows literally everything there is to know about the perception and the experience of colour, but because she has been locked up in a dimly-lit, black-and-white room her entire life, she never saw colours herself. Then, shortly after her twenty-third birthday, she is exposed to the bright world of red, yellow and blue. Question: what happens? Does Mary, in actually seeing colours, learn anything she did not know before? Of course she does!, answers Jackson (Jackson 1982). Of course she does not!, counters Paul Churchland (Churchland 1985). Who is right? Again, there is no way to tell. This is not to say that Jackson and Churchland have not given arguments for their beliefs. They have, but only to shunt the disagreement to another level. Thus Churchland chides Jackson for having failed to distinguish between knowledge by description and knowledge by acquaintance, so that wrongfully it looks as though Mary, after her liberation, comes to know something she did not know before in the same sense of 'to know'. Jackson, in turn, argues that the distinction between knowledge by description or acquaintance is not germane to the question, which is about *what* Mary knows and not about *how* she knows it (by description or by acquaintance) (Jackson 1986). This debate is doomed continually to go round on a merry-go-round, since the participants more or less tacitly endorse two totally different starting points. Churchland believes that, since what Mary learns after her release is not propositional knowledge, it fails to be relevant for an explanation of consciousness. But Jackson believes that, since Mary did learn something new, it *is* relevant for such an explanation. But of course this controversy cannot be resolved, since there is, at present, no way to know whether what Mary learns is or is not relevant for explaining consciousness. And as long as we do not know what *is* relevant, the same Mary-story can be happily taken in two opposite ways.

3. SECOND INDICATION: CONCLUSIONS BEG THE QUESTION

The second clue is connected to the first, indeed it takes the first indication one step further. It is that 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. In other words, not only are the conclusions contradictory, they also include the intuitions for the sake of whose elucidation the thought experiment was constructed.

The thought experiments about physical replicas may again serve as example. Not only do we have here two conclusions that contradict one another, these conclusions also embody the intuitions for which the entire thought experiment was set up in the first place. For thought experiments about physical *Doppelgänger* are meant to assess our intuitions about the mental as distinguished from our intuitions about the physical. These preliminary intuitions should therefore not appear in the final conclusions.

Searle's Chinese Room experiment is another famous example of begging the question. Several of Searle's opponents have remarked that the man in the room (in the original version Searle himself) is only a subsystem. A subsystem can of course not be said to understand Chinese, any more than can a lobe in my brain or a central processing unit in a computer be taken to have this ability. What does understand Chinese, Searle's opponents argue, is the entire system: the room plus the baskets with the Chinese symbols plus the book of instructions, et cetera.

In answer to this criticism, Searle invented a new version of his experiment, one in which the man in the room no longer needs the instruction book and the baskets, because he has learned the book by heart and can draw the Chinese symbols himself. Searle is now justified in stating that the man himself, and nothing more, forms the entire system. At the same time, however, he has less reason to maintain that the man does not understand

Chinese but is only manipulating symbols. To be sure, Searle sticks to his guns: 'They [the opponents] argue that it is the whole system ... that understands Chinese. But this is subject to exactly the same objection I made before. There is no way that the system can get from syntax to semantics. I, as the central processing unit have no way of figuring out what any of these symbols means; but then neither does the whole system'. (Searle 1984, 34). It will be clear that others can now easily claim the opposite, and they have done so in every possible way. Our claim is that, at this point, the statements pro and con beg the question: they reiterate those intuitions about 'understanding' and 'knowing what a symbol means' that were meant to be adjudicated by the very thought experiment in question.

At this juncture, one might think that philosophical thought experiments alone suffer from the above weaknesses. Thought experiments in science, one might think, are free from these blemishes. Our example of Galileo's successful thought experiment might well have suggested as much (but see Atkinson and Peijnenburg, forthcoming). Moreover, the idea seems to be fostered by Kathleen Wilkes' interesting analysis, implying that, due to the difficulty of deciding which parameters are relevant and which are not, in practice nearly all thought experiments in philosophy are poor ones, while those in science are not (Wilkes 1988).

Nevertheless, in science, too, defective thought experiments occur. We shall first give an example of a scientific thought experiment that generates contradictory conclusions (Section 4), and then one of a thought experiment for which the conclusions beg the question (Section 5). However, although the thought experiments in themselves are and remain poor ones, the scientific setting is such that the consequences are not as catastrophic as they are in the philosophical case.

4. NEWTON'S BUCKET AND EINSTEIN'S SPHEROID

An example of a scientific thought experiment that produces contradictory conclusions is the thought experiment that is known as Newton's bucket, but that might well have been called Einstein's spheroid. For simplicity of exposition we first describe Einstein's spheroid, then Newton's bucket, and then we explain why they are illustrations of essentially the same phenomenon. Finally, we show why the bucket/spheroid is just as suspect as the two philosophical thought experiments of Section 2.

In 1916, in 'The Foundation of the General Theory of Relativity', Einstein described the spheroid experiment as follows:

> 'Two fluid bodies of the same size and nature hover freely in space at so great a distance from each other and from all other masses that only those gravitational forces need be taken

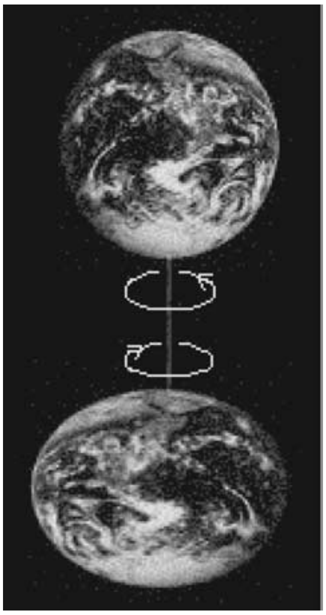

Figure 1.

> into account which arise from the interaction of different parts of the same body. Let the distance between the two bodies be invariable, and in neither of the bodies let there be any relative movements of the parts with respect to one another. But let either mass, as judged by an observer at rest relatively to the other mass, rotate with constant angular velocity about the line joining the masses. This is a verifiable relative motion of the two bodies. Now let us imagine that each of the bodies has been surveyed by means of measuring instruments at rest relatively to itself, and let the surface of S1 prove to be a sphere, and that of S2 an ellipsoid of revolution. Thereupon we put the question - What is the reason for this difference in the two bodies?' (Einstein 1916, 112).

Here Einstein asks us to imagine two separate fluid bodies, S1 and S2, rotating with respect to one another around the virtual line that connects their centres (see Figure 1). Since by assumption their relative motions are the same (albeit in opposite directions), one might expect S1 and S2 to have the same shape as well. Then how is it possible that S1 is a sphere, as measured by a man on its surface, whereas S2 is an ellipsoid for a woman on the surface of S2?

Two and a half centuries earlier, in the Scholium after the Definitions at the beginning of the *Principia*, Newton described an experiment known as Newton's bucket. It involves 'a vessel, hung by a long cord, that is so often turned about that the cord is strongly twisted, then filled with water' (Newton 1686, 10). Newton goes on to explain that, before the vessel is released, the vessel and the water in it are at rest, and the surface of the water is plane (i.e. flat). But when the vessel is released, it begins to rotate, and soon after the water in it rotates too, its surface becoming concave. Why is it that the water surface is plane at first and concave later?

Newton's thought experiment is in the relevant aspect equivalent to Einstein's. Where Einstein imagines one spherical fluid body S1 and one ellipsoidal body S2, Newton observes a plane water surface at the beginning of the experiment and a concave water surface at the end. Furthermore, where Einstein asks what it is that causes the difference between S1 and S2, Newton asks what it is that causes the difference between the surface being flat and concave.[3]

Newton's answer to both questions would have been to invoke absolute space. For him, S1 is a spheroid because it is at rest (or in uniform motion) with respect to absolute space, whereas S2 is an ellipsoid because it rotates with respect to absolute space. Similarly, Newton states that at the beginning of the bucket experiment the water is (approximately) at rest relative to absolute space, and at the end it rotates relative to it.

Einstein will have none of this. He does not accept the existence of absolute space, let alone that it could exert causal power. He castigates Newton's answer on the grounds that it invokes a wholly unacceptable '*factitious* cause' (Einstein 1916, 113).[4]

As to the question about what causes the difference between S1 and S2, he says: 'No answer can be admitted as epistemologically satisfactory, unless the reason given is *an observable fact of existence.* ... Newtonian mechanics does not give a satisfactory answer to this question'. (Einstein 1916, 112-113, author's italics). Then how does Einstein explain the bucket/spheroid? The answer is already indicated by the quotation above. The cause of both the difference between S1 and S2 and that between the flat and concave surface, must be 'an observable fact of experience'. Since space is not such an observable fact, but actually nothing but the separation between bodies, it cannot function as a cause. For Einstein, '[t]he law of causality has not the significance of a statement as to the world of experience, except when *observable facts* ultimately appear as causes and effects" (*ibid.*, 113; italics by the author). Hence for him a cause must be a material thing, i.e. another body. Taking his inspiration from Mach, for whom the inertial mass is determined by the distribution of matter in the rest of the universe, Einstein argued that the difference between S1 and S2 – and, similarly, the difference between the flat and the concave water surface – is caused by the influence of matter in the rest of the universe. Thus the man on S1 sees that the stars above his head are fixed, and this explains why his own fluid planet is a sphere. He reasons as follows: 'I can see (fixed) stars, and Mach tells me that my planet S1 is influenced by them. This influence must be symmetric in all directions, for only that can explain why, according to my own measurements, S1 is a sphere'. On the other hand, the woman on S2 sees the stars moving from East to West, and for her this explains the ellipsoidal shape of her fluid planet. Her reasoning is: 'From Mach I have learned that the celestial bodies that I see affect my planet S2. Apparently this effect is different at the equator than it is at the poles, for according to my measurements, S2 is an ellipsoid'. If the man and the woman had been sitting on the edge of Newton's bucket, their feet dangling down into flat and concave water respectively, their reports on the heavenly bodies and their reasonings would have been analogous.

Why can the bucket/spheroid thought experiment be put on a par with the philosophical thought experiments that we described in Section 1? The answer is: here, too, diametrally opposed conclusions have been drawn by different people from the same experiment. On the basis of the bucket experiment, Newton concludes that absolute space exists, whereas Einstein, on the basis of essentially the same experiment, claims that this is not the case.

But although this answer is true, it is not very illuminating. What remains to be shown is that, at the level of this thought experiment, *it cannot be decided* whether Einstein or Newton is right. After all, in the philosophical thought experiments discussed in Section 1, the key observation is that there is *no reason* to prefer the one conclusion over the other. Or, as Kant says (*vide infra*), the problem of the conflicting conclusions is that 'it is impossible to decide between them' (A501, B529). It is exactly this undecidability that threatens the use of thought experiments as a means to settle a debate.

The undecidability in the case of the bucket/spheroid is shown best by focusing on the fact that neither Newton nor Einstein are able to justify the principles they invoke. Newton 'explains' the happenings in the bucket experiment on the basis of the rotation or non-rotation with respect to absolute space; but the only justification for the existence of absolute space is that it can explain happenings such as those in the bucket experiment. Newton's theory can be stated perfectly well without the concept of absolute space: we do not need absolute space to give meaning to his first law of motion (viz., a body continues in its state of rest, or of uniform motion in a straight line, unless it is acted on by forces). It is sufficient to postulate the existence of a class of reference systems (nowadays called Galilean systems), each with a constant velocity relative to the others, with respect to which all three of Newton's laws hold. No one of these reference systems plays a special role, and the notion of absolute space is neither invoked nor needed.

Einstein recognized that Newton's principle of absolute space was empty. However, what he did not see, at least not at the time that he wrote his 1916 paper, was that the same goes for the principle he himself invokes. Einstein never succeeded in justifying Mach's Principle, which maintains that the rest of the universe gives rise to the shapes of S1 and S2.[5] Indeed, Einstein did not implement Mach's Principle in his general theory of relativity: it always remained a mere decoration without empirical content, to be superadded or removed in accordance with the theoretician's taste. A year before his death, Einstein admitted as much: 'As a matter of fact, one should no longer speak of Mach's principle at all'.[6]

However, the fact that contradictory conclusions are drawn from one and the same thought experiment is here by no means as disastrous as it was in the analogous philosophical case. This becomes clear as soon as we leave the level of the thought experiment itself and turn to the theories in question. For when we step over the boundaries of the bucket/spheroid and look at the theories they inspired, we *do* have some reason to prefer the one theory over the other. The theory we prefer is, of course, Einstein's. For although Newton built an empirically successful system of mechanics that held undisputed sway for three and a half centuries, Einstein's mechanics was even better. By denying all privileged status of one system of reference coordinates above another, Einstein took an Olympian standpoint from which he could build his general theory of relativity. This theory defines the domain of validity of Newton's theory (namely when all gravitational effects are weak and all speeds are small compared with that of light), and it effectively replaces the Newtonian concept of gravitational force by that of the geometry of space-time.

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? It looks as though, at the moment, we have no more than thought experiments here, and these thought experiments leave much to be desired.

In the next section we will see that scientific thought experiments, in addition to generating contradictory conclusions, can also beg the question. But again the consequences are less dramatic, because again we will have something else at our disposal, this time a real experiment.

5. EINSTEIN, PODOLSKY AND ROSEN

Against Bohr and other representatives of the Copenhagen school, who claimed that quantum mechanics can describe every phenomenon in its domain, Einstein, Podolsky and Rosen (EPR) have argued that quantum mechanics is incomplete. Quantum mechanics, conceived as a theory of phenomena at the atomic level, is in the view of EPR unable to account for all those phenomena. In particular, atomic positions and velocities exist that quantum mechanics cannot describe.

It has been recognized for a long time that measuring an object involves exchanging energy between the object and the measuring apparatus. According to quantum mechanics, energy is not infinitely divisible, but exists in small packets (quanta) of definite size. As a consequence, measuring a property of a particle involves the exchange of at least one quantum, and hence implies a non-negligible disturbance of the particle. For example, if one measures the velocity of a particle at a certain time t, one cannot measure its position at t. At best the position can be measured briefly thereafter, but the particle has then been so disturbed that one can no longer reconstruct its position at t. This makes it problematic to maintain that both the velocity and the position have definite values at t. For this reason indeed, Niels Bohr denied that both the velocity and the position at t exist.

In an attempt to show that Bohr's position is wrong, and that both the velocity and the position exist even if we cannot measure them both, EPR constructed the following thought experiment. Suppose that two particles are created together in such a way that their positions and their velocities at later times remain correlated. That is, if one measures the position of particle 1 at one of those later times, t, one can calculate, and thereby predict with certainty, the position of particle 2 at t. A similar story can be told for the velocities: if at t one measures the velocity of particle 1, one can immediately infer the velocity of particle 2. Although one would have to decide whether to measure the position or the velocity of particle 1 at t – one could not do both – EPR stress that this decision does not affect particle 2. After all, the two particles could be miles or even light years apart when the measurement on particle 1 is made, so that, although the measurement will disturb particle 1 and change its properties, no such disturbance could be instantaneously transmitted to particle 2 and change *its* properties. Since either property could have been measured for particle 1 and thus be predicted with certainty for particle 2, and since the ontological status of particle 2 does not depend on whether one decides to measure the position or the velocity of particle 1, it follows that the position *and* the velocity of particle 2 must exist:

'If, without in any way disturbing the system, we can predict with certainty (i.e., with probability equal to unity) the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity'. (Einstein, Podolsky and Rosen 1935, 777).

According to Bohr, the position and the velocity of particle 2 cannot exist simultaneously any more than can the position and the velocity of particle 1. For a phenomenon only exists when it has actually been measured. Einstein, on the other hand, claims that a phenomenon exists if it is predicted with certainty. In other words, Einstein ties physical reality to absolutely correct *predictions* that *could* be made; Bohr ties it to *measurements* that *actually are* made.

We have here a thought experiment with contradictory conclusions: quantum mechanics is complete versus quantum mechanics is not complete, or something exists when you have *in fact* measured it versus something exists when you can infer it *in principle*. Moreover, these conclusions beg the question, for they are embodiments of those intuitions for the sake of which the entire thought experiment was conceived. What was at stake at the beginning of the debate was precisely the question what is or is not an element of physical reality, and it is inappropriate to present those initial intuitions as final conclusions.

However, although this thought experiment failed (as a thought experiment), the consequences were again not as disastrous as they were in the case of the philosophical thought experiments. In 1952 David Bohm retooled the EPR experiment in such a way that John Bell, in 1964, could make a prediction. This prediction, the so-called Bell inequality, was based on the assumption that local hidden variables exist and thus that quantum mechanics is not complete. It was tested in 1982 by Alain Aspect, and the result appeared to be negative. Hence hidden variables do not exist, quantum mechanics is complete, and Bohr *cum suis* were right. Although we do not want to suggest that the last word has been said, it is true that most physicists and philosophers have deserted Einstein and rallied to Bohr's colours.

The EPR-experiment has thus been given a testable format, but it is unclear how we ever could put the Chinese Room or the Mary experiment to the test. To be sure, both the Chinese Room and the Mary experiment can be carried out, ethical considerations aside, but that would not resolve the philosophical conundrum.[7]

6. BETWEEN THEOLOGY AND SCIENCE

We have elucidated two grounds for suspecting a particular thought experiment to be a poor one. These grounds are relevant to philosophical and scientific thought experiments alike. However, our examples suggest that in science the damage is restricted: in the bucket/spheroid example we can still fall back on theories, whereas the EPR-thought experiment finally gave rise to a real experiment, and an *experimentum crucis* at that. In philosophy such escape routes generally do not exist.

We are not trying to say that philosophy should resemble science. Philosophy is and should be different from science; in particular it is and should be more speculative. (This is true even if philosophy is deemed to be continuous with science.) But one can endorse this difference, and still take two entirely different attitudes towards philosophical thought experiments. The one attitude is to say that philosophical thought experiments are, and should be, more speculative than their scientific cousins. On this view, the production of fancy thought experiments is fine and should go on – it suits the philosophical genre. The other attitude takes it that, since philosophy is speculative by its very nature, one should not make it more speculative by concocting recherché thought experiments. On the contrary, one should try to find an antidote: try to make philosophy more empirical, for instance.

We tend on the whole to the latter option. It was Bertrand Russell who, on the first page of his *History of Western Philosophy*, characterised philosophy as 'something intermediate between theology and science'. We fear that the prevailing outré thought experiments pull analytic philosophy in the direction of theology (it seems no coincidence that omniscient beings feature prominently in some famous thought experiments). In the same vein, it was pulled towards science in the early days of the logical positivists.

But rather than being pulled in a certain direction, philosophy should keep its intermediate position. To do so, it needs a corrective. Being by nature a speculative enterprise, philosophy benefits from non-speculative input, such as empirical facts and theories. Science, on the other hand, being testable and less speculative, seems to benefit from speculations such as thought experiments. Just as strict operational definitions are often advantageous in the social sciences whereas they are obstructive in physics, so thought experiments seem to fare better in natural science than in philosophy.

7. KANT'S ANTINOMIES OF PURE REASON

Our exposé might have reminded one of Kant. Especially the philosophical examples mentioned in Section 2 bear a striking resemblance to the metaphysical sophistries that in the *Critique of Pure Reason* are described as 'antinomies' or 'dialectical oppositions'. Whether the world has a beginning in time and a limit in space; whether there exists within us an indestructible unity; whether we are free or are bound by the chains of nature and of fate; whether there is a supreme cause of the world or whether all our thought and speculation must end with nature: these are questions that we, with our cognitive equipment, are unable to resolve. For there is no object of experience that corresponds to the ideas expressed in these questions. Analogously, there are no concepts with which these ideas can be described: such concepts lie outside the sphere of possible experience that is necessary for them to be meaningful and coherent. If we ignore these facts and do try to settle the questions, we will end up with sophisticated arguments *pro* and *con*, which show that conclusion *X* as well as *not-X* can be proved. In this situation, the reasonable thing to do is to go back to the source of the debate, where we will discover that the entire discussion is built on sand:

> '... reason, in the midst of its highest expectations, finds itself so compromised by the conflict of opposing arguments, that . . . nothing remains for reason save to consider whether the origins of this conflict, whereby it is divided against itself, may not have arisen from a mere misunderstanding'. (A464–465, B492–493).

Kant's antinomies draw the boundaries of the logical system as it was then conceived, being based on Aristotle's fourfold classification of general statements: All S are P, No S is P, Some S's are P, Some S's are not P. The crucial terms in the antinomies – 'world', 'immortal soul', 'God', 'freedom' – are such that the subject-predicate statements in which they occur do not respect the boundaries of the Aristotelian logic of nonempty classes (De Jong 2000). Whenever we try to reason with statements that have one of these terms as subject term, we will find that the relations of contradiction, contrarity, subalternation etc. do not hold. From this Kant concluded that any discussion about the truth or falsity of those statements is nugatory:

> '... nothing seems to be clearer than that since one of them asserts that the world has a beginning and the other that it has no beginning ... one of the two must be in the right. But even if this be so, none the less, since the arguments on both sides are equally clear, it is impossible to decide between them. The parties ... are really quarrelling about nothing, and ... a certain transcendental illusion has mocked them with a reality where none is to be found'. (A501, B529).

Were we to take Kant's lessons to heart, we would conclude from the two thought experiments in Section 2 that they cannot teach us anything about consciousness, quale, or personal identity. Just as concepts like 'world', 'soul', and 'God' evidently step over the boundaries of standard Aristotelian logic, so concepts like 'consciousness', 'quale' and 'personal identity' ostensibly escape the boundaries of our thought experiments. Just as 'having a beginning' or 'not having a beginning' apparently are not predicates that can be applied to 'world' as a subject term, so 'being physical' or 'not being physical' apparently are not predicates that can be applied to a subject term like ‘consciousness’. Any context in which these predicates *are* connected to these subject terms can generate conflicting conclusions.

In the light of our claim that the occurrence of contradictory conclusions seems less disastrous in science than in philosophy, it is interesting to recall what happened to Kant’s four antinomies of pure reason. The three more philosophical antinomies (on the alleged existence of God, free will, and the immortality of the soul) are still with us, and may well be undecidable, forever saddling us with contradictory conclusions. However, the one scientific antinomy, on cosmology, has long been relegated to the status of a detail of non-Euclidean space-time geometry. In the standard big-bang theory, the universe is finite in present spatial extent and in past temporal duration. The latest observations, however, seem to suggest that the universal expansion will not be reversed in the future, which implies that future duration will have no limit. Since the expansion is not one of matter into pre-existing space, but rather one of matter and energy that engender non-Euclidean space-time, the puzzle as to how the universe can be spatially finite but unbounded can at least be apprehended mathematically. Whether the universe has a finite or an infinite future is perhaps still a matter of debate, but this is no longer a philosophical puzzle, let alone an antinomy. The basic question is whether the average mass-density of the observable universe is sufficiently high to lead to reversal of the expansion, leading to a *big crunch* at a future, finite time, or not. The competing theories answering that question make distinct, testable predictions.[8]

NOTES

[1] Alexander Bird and Richard Holton have remarked that a modal rendition of this formula might well be preferable (private conversation). Holton suggested $(T \rightarrow \Box S)$, $(E \rightarrow \neg S)$, $\Diamond E \vdash \neg T$, whereas Bird proposes as the simplest modification $(T \,\&\, \Diamond E) \rightarrow \Diamond S$, $\neg \Diamond S$, $\Diamond E \vdash \neg T$ Another account, also according to Bird, would be $T \rightarrow \Box (E \rightarrow S)$, $\Diamond (E \,\&\, \neg S) \vdash \neg T$. Holton and Bird have a point, but the simple formula suffices for our purposes. More on modalities in destructive thought experiments: Sorensen 1992, Chapter 6, Section II.

[2] John Norton has stressed that the particulars are useful, but not necessary for the point to be made. *Cf.* Norton’s definition of thought experiments as arguments that satisfy two conditions: they must (1) posit hypothetical or counterfactual states of affairs and (2) invoke particulars that are irrelevant to the generality of the conclusion (Norton 1991, 129).

[3] Actually Newton really did his experiment, for he talks about the water as ‘forming itself into a concave figure (as I have experienced) ... ’ *(ibid.)* whereas Einstein certainly could not perform his!

[4] The original German text says: ‘Der berechtigte Galileische Raum . . . ist aber eine *bloß fingierte* Ursache, keine beobachtbare Sache’.

[5] Mach’s Principle states more formally that inertial mass is engendered by the matter in the

rest of the universe. If the rest of the universe were removed, both of Einstein's spheroids would be perfect spheres and neither would have any inertial mass (according to Mach and Einstein, but entirely against the intuition of Newton, for whom absolute space would still exist, implying for him that S2 would remain ellipsoidal). Mach's thought experiment is breathtaking in its audacity, removing at a flourish the starry firmament above us; but it is also sterile so long as no theory exists that tells us how to calculate the difference in the inertial mass of, say, the earth as we know it, and what it would be if we were to remove the universe beyond our own galaxy, or perhaps beyond the cluster of which the Milky Way is a part.

[6] 'Von dem Mach'schen Prinzip sollte man eigentlich überhaupt nicht mehr sprechen'. In a letter to F. Pirani, February 2nd, 1954, quoted in Pais 1982, 288.

[7] This has been shown by Marjolein Degenaar for another thought experiment, viz. Molyneux's problem, named after the Irish philosopher William Molyneux, who presented it in 1688 to John Locke (Degenaar 1996). Imagine a man, blind from birth and capable of distinguishing and naming a globe and a cube by touch. Suppose that this man suddenly acquires sight. Would he be able to distinguish and name both objects simply by looking at them? The question provoked passionate discussions between 18th century rationalists and empiricists, who answered it with *yes* and *no* respectively. As Degenaar notes in her excellent book on the subject, the discussion took a new turn after the first cataract operations. From a pure thought experiment, Molyneux's problem turned into a question that could be answered on the basis of real experimentations. Surprisingly enough, however, no agreement was reached. Rationalistic and empiristic scholars kept harassing each other, now over the correct interpretation of the experiment. Degenaar concludes that Molyneux' problem cannot be solved empirically. If she is right, then the Chinese Room and the Mary experiment can certainly not be resolved.

[8] We would like to thank audiences in Leusden (International School for Philosophy, The Netherlands), Mexico City (Department of Philosophy, UNAM) and Edinburgh (Department of Philosophy, University of Edinburgh) for their valuable remarks.

REFERENCES

Atkinson, D. and Peijnenburg, J.: forthcoming, 'Galileo and Prior Philosophy', *Studies in History and Philosophy of Science.*

Bishop, M. A.: 1999, 'Why Thought Experiments are not Arguments', *Philosophy of Science* **66**, 534–541.

Block, N. *et al.*: 1997, *The Nature of Consciousness*, Cambridge, Mass.-London, the MIT Press.

Borsboom, D., Mellenbergh, G. J. and van Heerden, J.: 2002, 'Functional Thought Experiments', *Synthese* **130**(3), 379–387.

Brown, J. R.: 1991, *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*, London-New York, Routledge.

Chalmers, D. J.: 1996, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford, Oxford U.P.

Churchland, P.: 1985, 'Reduction, Qualia, and the Direct Introspection of Brain States', *Journal of Philosophy* **82**, 8–28.

Churchland, P.: 1989, 'Knowing Qualia: A Reply to Jackson', reprinted in: Block 1997.

Degenaar, M.: 1996, *Molyneux's Problem: Three Centuries of Discussion on the Perception of Forms*, Dordrecht, Kluwer.

Dennett, D. C.: 1997, ‘An exchange with Daniel Dennett’, in: J. R. Searle, *The Mystery of Consciousness*, New York: A New York Review Book.

Einstein, A.: 1916/1923, ‘The Foundation of the General Theory of Relativity’, in H. A. Lorentz *et al.*, *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity*, with notes by A. Sommerfeld. Transl. by W. Perret and G.B. Jeffery (New York, Dover Publications).

Einstein, A., Podolsky, B. and Rosen, N.: 1935, ‘Can Quantum-Mechanical Description of Physical Reality be Considered Complete?’, *Physical Review* **47**, 777–780.

Gendler, T. S.: 1998, ‘Galileo and the Indispensability of Scientific Thought Experiment’, *British Journal for the Philosophy of Science* **49**, 397–424.

Hofstadter, D. R.: 1981, ‘Reflections on Searle’s “Minds, Brains, and Programs’ ”, in D. C. Dennett and D. R. Hofstadter (eds), *The Mind’s I: Fantasies and Reflections on Self and Soul*, New York, Basic Books.

Horowitz, T. and Massey, G. J. (eds): 1991, *Thought Experiments in Science and Philosophy*, Savage, Maryland, Rowman and Littlefield.

Jackson, F.: 1982, ‘Epiphenomenal Qualia’, *Philosophical Quarterly* **32**, 127–136.

Jackson, F.: 1986, ‘What Mary didn’t Know’, reprinted in: BLOCK 1997.

Jong, W. R. de: 2000, ‘Kants Dialectische Opposities (Kant’s Dialectical Oppositions)’, *Algemeen Nederlands Tijdschrift voor Wijsbegeerte (General Dutch Journal of Philosophy)* **92**(2), 154–160.

Kant, I.: 1781–1787/1929, *Critique of Pure Reason*, Hampshire-London, Macmillan Press, Transl. by N. Kemp Smith, 25th edition, 1993.

Kim, J.: 1996, *Philosophy of Mind*, Boulder-Oxford, Westview Press.

McAllister, J. W.: 1996, ‘The Evidential Significance of Thought Experiment in Science’, *Studies in History and Philosophy of Science* **27**(2), 233–250.

Newton, I.: 1686–1626/1729/1934, *Philosophiæ Naturalis Principia Mathematica*, Andrew Motte’s 1729 translation into English, revised by Florian Cajori (Berkeley etc., University of California Press). Paperback edition 1966.

Norton, J. D.: 1991, *Thought Experiments in Einstein’s Work*, in: Horowitz and Massey.

Norton, J. D.: 1996, ‘Are Thought Experiments just what you Thought?’, *Canadian Journal of Philosophy* **26**(3), 333–366.

Parfit, D.: 1984, *Reasons and Persons*, Oxford, Oxford U.P. 5th reprint with corrections, 1991.

Pais, A.: 1982, *Subtle is the Lord … The Science and the Life of Albert Einstein*, Oxford, Oxford U.P.

Searle, J. R.: 1980, ‘Minds, Brains, and Programs’, reprinted in D. C. Dennett and D. R. Hofstadter, *The Mind’s I: Fantasies and Reflections on Self and Soul*, New York, Basic Books.

Searle, J. R.: 1984, *Minds, Brains and Science*, reprinted in 1989, London etc: Pelican Books.

Sorensen, R.: 1992, *Thought Experiments*, Oxford, Oxford U.P.

Wilkes, K.: 1988, *Real People: Personal Identity Without Thought Experiments*, Oxford, Oxford U.P. Paperback Edition 1993.

Faculty of Philosophy
University of Groningen
Aweg 30
9718 CW Groningen
The Netherlands
(peijnenburg@philos.rug.nl)