When are Discussions of Thought Experiments Poor Ones?[1]

A Comment on Peijnenburg and Atkinson

Daniel Cohnitz (Düsseldorf)

Abstract

In their recent paper, "When are thought experiments poor ones?" (Peijnenburg/Atkinson 2003), Jeanne Peijnenburg and David Atkinson present an argument to the conclusion that most, if not all, philosophical thought experiments are "poor" ones with "disastrous consequences" and that they share this property with some (but not all) scientific thought experiments. The moral they draw is that the use of thought experiments in science is generally more successful than in philosophy (of mind).

In this comment I shall briefly try to show that Peijnenburg's and Atkinson's view on thought experiments as it is presented in Peijnenburg/Atkinson 2003, but also in Atkinson/Peijnenburg [forthcoming], and Atkinson 2003, is based on an misleading characterization of both, the dialectical situation in philosophy as well as the history of physics. By giving an adequate account of what the discussion in contemporary philosophy is about, we will arrive at a quite different evaluation of philosophical thought experiments.

I. Peijnenburg's and Atkinson's view on thought experiments

According to Peijnenburg and Atkinson ('PA', for short) there is a felt unease with many philosophical thought experiments. Indeed, during the last twenty years there has been a considerable amount of papers and monographs on the matter, whereas most contributions were rather critical and suggested either to replace thought experimentation in philosophy with other methods, or to set up methodological standards that are to be met by successful thought experiments.[2]

On the other hand there have been also quite a few defenders in recent years. Frank Jackson, David Chalmers, Stephen Yablo, George Bealer, John Perry have argued (on very different grounds) that thought experiments are quite useful tools of philosophical research.[3]

PA refuse to give a definition of what they take 'thought experiment' to refer to, since they do not think that this would be necessary for an account of whether or when a thought experiment is a good one. Of course, that is false. The word 'thought experiment' has a rather

[1] I would like to thank Axel Bühler, Mark Breuer, David Chalmers, Klaus-Jürgen Düsberg, and Marcus Rossberg for helpful comments on earlier versions of this paper.

[2] For an overview see Cohnitz 2003c.

[3] Jackson 1998, Chalmers 2002, Yablo 1993, Bealer 1998, Perry 2002, see also Cohnitz 2003b,c.

broad meaning and includes, as Elke Brendel[4] and others have pointed out, also imaginary cases described in physics textbooks. These thought experiments are usually meant to help students of physics understand the implications of certain theories and should be judged in relation to how well they accomplish that task. The unrestricted notion also includes thought experiments that are not intended to induce *justified* belief revision in the addressee, but belief revision that is achieved by ridiculing a rival theory or by recommending the own (as devices of propaganda). These possible uses of thought experiments have special criteria of adequacy, most of which will be the subject of empirical psychology or the sociology of science.

However, as it seems clear from their discussion, PA are interested in a subclass of thought experiments only. The subclass they are interested in is what we might call "critical thought experiments" or "necessity refuters" or "destructive thought experiments". Such thought experiments can be characterized as follows[5]:

(G1) It's the aim of a destructive thought experiment to induce justified belief revision in the addressee.

(G2) To achieve this aim, a certain state of affairs is described which is claimed to be conceivable.

(G3) The aim is intended to be achieved without realizing the described state of affairs or assuming that the state of affairs really obtains.

(G4) If T is the belief in the addressee the destructive thought experiment is intended to revise, S a statement whose necessity follows from T, E the state of affairs claimed to be conceivable and therefore possible, the logical structure of a thought experiment can in general be regimented thus: $(T\rightarrow\Box S)$, $(E\rightarrow\neg S)$, $\Diamond E \vdash \neg T$.[6]

I.1 *Atkinsons and Peijnenburg on Philosophy*

PA identify a number of philosophical thought experiments which obey the characterization given. The examples discussed are Searle's Chinese Room, Frank Jackson's Mary the color scientist, and Chalmers' zombie. All these thought experiments are "poor ones" according to PA, for they allegedly satisfy at least one of the following indications for a poor thought

[4] Brendel 1999.
[5] (G1)-(G3) are conditions first defined by Ulrich Gähde, in Gähde 2000. For a discussion see Cohnitz 2003c.
[6] This regimentation differs from Atkinson's and Peijnenburg's and follows Cohnitz 2003b. Atkinsons and Peijnenburg seem to admit that a regimentation like this is more adequate than theirs.

experiment: (i) their "conclusions contradict one another" or (ii) their "conclusions beg the question". We will briefly try to explicate what is meant by these two indications.

(i) conclusions contradict one another

Well, certainly no philosopher of modern analytic philosophy has managed to present an argument with two contradicting conclusions. It also seems that "conclusion" – as intended by PA – does not refer to the states of affairs necessitated by the target theory T, such that "conclusions" would "contradict" if $(T \rightarrow (E \rightarrow (S \wedge \neg S)))$. For something like this seems to be going on in Galileo's famous Pisa experiment (at least according to the reconstruction given in Peijnenburg/Atkinson 2003), from the Aristotelian theory (T) of falling bodies, two mutually exclusive predictions follow if applied to the case (E) that a lighter musket shot and a cannon ball fall connected with each other by a rod. The Aristotelian theory of falling bodies would predict for this case that the compounded system both falls faster *and* slower than the musket shot alone. These are "contradicting conclusions" but not of the kind PA seem to be having in mind.

What they seem to be having in mind is that confronted with the description of a scenario, no clearly shared intuitions are triggered in the audience. If asked whether zombies are conceivable, or whether Mary learned something new, part of the audience would agree the other disagree, presumably, in accordance with their antecedent convictions. Moreover, these thought experiments are said to lead to "disastrous consequences", because both sides not only *can* but *do* use the same thought experiment to argue for their position. The only thing changing is the intuition. Thus the indication is that a thought experiment is a poor one if the described scenario does not trigger a uniform intuition in the audience, but intuitions that oppose each other exactly concerning the topic at issue.

(ii.) conclusions beg the question

To speak of "question begging conclusions" in an accusing tone seems to involve a too broad notion of "question begging". Isn't it kind of natural for a *conclusion* to beg the question? Conclusions of philosophical arguments are intended to answer philosophical problems, thus – of course – they should take some definite stance. "'Physicalism is false' is the conclusion of your argument, but whether physicalism is true or false is what is at issue in our discussion,

thus you are begging the question." is not a correct application of a question begging-accusation.[7]

PA had better meant something different. It is not entirely clear from their paper what they do mean, but here is a less controversial example that might clarify. Consider the following thought experiment by Gottfried Wilhelm Leibniz:

> Let us assume that some individual should instantly become the King of China, with the further constraint that he would lose all memory concerning his former life just as if he would be born anew. Wouldn't that – practically speaking – be indistinguishable from the situation in which the individual would be destroyed and a King of China brought into being at the same moment and the same place? The individual has no reason to wish for that. (Leibniz 1686/1996, 145-157)

Thus Leibniz himself suggests that it would amount to a content less incoherent wish if one wanted to be the King of China by thereby losing the memory of his own past. Here is the same story with a different moral drawn:

> Many people wish they were somebody else in the sense that they wish they were in his shoes with his body, position, relationships, appearance, memory and character. Perhaps my body is withered, my own position and relationships are unsatisfactory, my looks are ugly, my memories give me no joy, and I am profoundly dissatisfied with my own character. You on the other hand seem very satisfactory in these ways. So I wish I were you (in the above sense). Is the wish coherent? Am I, that is, wishing for the existence of a logically possible state of affairs different from the present state? Superficially, yes. (Swinburne 1974, 245)

Leibniz and Swinburne seem to have exactly opposite intuitions regarding this case. Which intuitions they have seems to be depending on what they believe personal identity to consist in. Leibniz seems to have endorsed – at this time of his philosophical career – a reductionistic analysis, pretty close to Locke's theory, whereas Swinburne thinks that personal identity is something unanalyzable.[8] If the intuitions triggered by philosophical thought experiments were by and large dependent on what *antecedent* convictions we have, they could not fulfill the critical role described above. They would still be useful, of course, for they could serve to illuminate what a theory says, or inform us about which theory we are implicitly holding.[9] I

[7] Nevertheless, Atkinson and Peijnenburg are talking that way when criticizing the EPR thought experiment: Atkinson/Peijnenburg 2003, 317.
[8] Later Leibniz came to hold the same view. See his *Nouveaux Essais Sur L'Entendement Humain*. For a more detailed analysis see Noonan 1983, 57-64.
[9] Some interesting applications are the Personal Identity Game and similar applets in the internet (http://www.philosophers.co.uk/games/identity.htm). The user is confronted with imaginary cases and has to

think, PA would agree that (given this presentation) the thought experiment above satisfies both indications for a "poor" thought experiment.

Unlike most other critics of philosophical thought experimentation, PA think that also some scientific thought experiments are "poor ones". The reason is that even some scientific thought experiments seem to have at least one of the indicating properties. Nevertheless, scientific thought experiments, even poor ones (as judged by the both indicators), are said to have less "disastrous" consequences (and might then turn out to be "very good ones", after all). Let's see what that means.

I.2 *Atkinson and Peijnenburg on Physics*

The two scientific thought experiments discussed are Newton's Bucket (or Einstein's spheroid) and the famous Einstein/Podolsky/Rosen thought experiment against the completeness of quantum mechanics.

Einstein's spheroid case from 'The Foundation of General Theory of Relativity'[10] was – according to PA – meant to show that observable effects have to be explained by reference to observable facts, not by an invocation of "absolute space". Whereas Newton's thought experiment, describing essentially the same case, was intended to prove the existence of absolute space. Einstein and Newton thus seem to draw contradictory lessons from one and the same thought experiment; the case seems to have indicating property (i.) and probably also (ii.). However, we know which account to favor. Einstein's theory is superior to Newton's, thus by turning to the theories supporting the contradicting conclusions, we arrive at a definite judgment. According to PA this is unlike the situation in philosophy where we allegedly have no reason to favor one account over the other:

> 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. (Peijnenburg/Atkinson 2003, 315)

The second example is EPR. Einstein, Podolsky, and Rosen designed a thought experiment to show that quantum mechanics is incomplete (that there are hidden variables, quantum mechanics does not account for). Again there seem to be contradictory conclusions drawn in the history of physics. Einstein et al. tied physical reality "to absolutely correct predictions

---

declare what he would say about these cases. In the end his answers are analyzed and he is informed about whether and which theory he implicitly holds or whether his answers were inconsistent.

[10] Einstein 1920.

that could be made"[11] and thus found quantum mechanics wanting. Bohr et al. tied it "to measurements that actually are made"[12] and considered quantum mechanics complete:

> We have 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, the 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. (Peijnenburg/Atkinson 2003, 317)

Although this thought experiment is by both indicators a poor one, it is considered by Atkinsons to be "a very good one"[13], for it led to an empirical test, namely the experiments by Alain Aspect in 1982. Although the original thought experiment left the situation indecisive, we could later turn to experiments which were inspired by the thought experiment. Again something that PA don't deem possible in philosophy:

> 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. (Peijnenburg/Atkinson 2003, 317)

I.3. *Summarizing Atkinson's and Peijnenburg's View*

PA seem to be endorsing the following theses:

(PA1) There are philosophical thought experiments (e.g., Zombies, Mary the color scientist) which trigger contradicting reactions (non-uniform judgments) in the audience which are endorsed by both sides of a philosophical dispute as arguments for the respective positions.

(PA2) There are philosophical thought experiments (e.g., Zombies, Searle's Chinese Room) which are question begging.

(PA3) Both of the above are true for physics as well (e.g., Newton/Einstein, EPR).

(PA4) In philosophy we cannot turn to theories to settle conflicts over thought experiments (whereas we can in science).

[11] Peijnenburg/Atkinson 2003, 317.
[12] Ibid.
[13] Atkinson 2003, 209.

(PA5) In philosophy we cannot turn to crucial experiments (whereas we can in science).

In the following I will throw doubt on all of them. Since it is philosophy which is under attack here, I will defend philosophical thought experiments (and thus address (PA1), (PA2), (PA4), and (PA5)) in section III. In the next section (II.) I will briefly turn to physics and the thought experiments mentioned, to discuss (PA3) and (PA5).

II. The Situation in Physics

PA claim that physics contains thought experiments that lead to contradicting and question begging conclusions. They also claim that in physics such disagreement can be solved by turning to crucial experiments. I will argue for the examples discussed by PA that the thought experiments in question do *not* lead to contradicting or question begging conclusions. There is disagreement, that is correct, but the disagreement is not based on "different intuitions" about what is happening in the counterfactual case considered, but based on conflicting criteria of adequacy for what counts as a satisfying explanation. This is fully independent of the method of thought experimentation, but a problem that occurs in all cases where we have to chose between competing explanations whose difference lies in the invocation of empirically indistinguishable theories.

II.1 *Buckets and Spheroids*

Newton's bucket experiment was intended to prove the existence of absolute space. First of all it is not really clear whether it really should count as a thought experiment. It rather seems to be a description of a real experiment that Newton had carried out, or a phenomenon one should be acquainted with (for the details of the phenomenon are certainly unlike what everybody would expect intuitively): A bucket is hanging from a long cord and turned as long as it takes to strongly twist the cord. Then the bucket is filled with water and held at rest. Suddenly the bucket is impelled in the opposite direction, while the cord untwists:

> […] the surface of the water will at first be flat, as before the bucket began to move; but after that, the bucket by gradually communicating its motion to the water, will make it begin sensibly to resolve, and recede little and little from the center, and ascend to the sides of the bucket, forming itself into a concave figure (as I have experienced), and the swifter the motion becomes, the higher will the water rise, till at last, performing its revolutions in the same times with the vessel, it becomes relatively at rest in it. (Newton 1726, 10; translation taken from Toretti 1999, 241)

Newton's explanation of the phenomenon is that the water's "endeavour to climb the bucket's walls and recede from its axis bears witness to its real rotation in absolute space, for the water's surface remained initially unchanged while it rotated only in appearance with respect to the adjacent bucket" (Torretti 1999, 241). Mach and Einstein are said to have come to the opposite conclusion by reflecting on the same thought experiment. Here is Mach's reading of the phenomenon:

> Newton's experiment with the rotating water bucket teaches us only that the rotation of water relative to the bucket walls does not stir any noticeable centrifugal forces; these are prompted, however, by its rotation relative to the mass of the earth and the other celestial bodies. Nobody can say how the experiment would turn out, both quantitatively and qualitatively, if the bucket walls became increasingly thicker and more massive – eventually several miles thick. (Mach 1883/1988, 256; translation taken from Torretti 1999, 241)

The disagreement over the "thought experiment" is what counts as the best explanation of an observable phenomenon. Newton's explanation assumes absolute space, whereas for Mach this is an unnecessary stipulation. The mass of the earth and the other celestial bodies are sufficient to explain the phenomenon *as it can be observed on earth*! (Thus Mach is critisizing Newton's methodology, Newton even seems to violate his own *rule of philosophy*, not to "allow more causes of natural things than are both true and sufficient to explain their phenomena" (Torretti 1999, 69).

So far there is no clash of intuitions concerning a thought experiment. The phenomenon is well known and observable, the discussion is about competing alternative explanatory hypotheses which are empirically equivalent.

There is a thought experiment mentioned in the last sentence of the Mach quote: what would happen if the bucket walls were several miles thick? Or, alternatively, what would happen if the water bucket would do his turns in an otherwise empty universe? Imagine we could observe centrifugal forces at a rotating body in an otherwise empty universe: let's imagine only one single planet would exist, wouldn't it be shaped like an ellipsoid if it rotated and a sphere if at rest? These imaginary cases would count as thought experiments. The latter case is a thought experiment by Carl Neumann (Neumann 1870).[14]

Do Mach and Newton disagree about these cases? For all cases Newton would certainly predict that the water in the bucket would show qualitatively the same behaviour and the lonely planet be thus shaped, for its motion is relative to absolute space. Would Mach, on

[14] Mach rejects it as inconclusive in Mach 1883/1988, 300.

the other hand, predict the opposite? No, Mach would rather refuse to answer the question. As becomes clear from the quote given above, we cannot trust our imagination in these counterfactual cases: "*Nobody can say* how the experiment would turn out, both quantitatively and qualitatively, if the bucket walls became increasingly thicker and more massive – eventually several miles thick" (my emphasis). Mach can even offer a theory of thought experimentation and physical intuition that would justify this move. We will not go into it here.[15]

Therefore in the discussion between Mach and Newton is no clash of intuitions so far. Both agree on what is going on in the water bucket on earth, but disagree about the proper explanation of the phenomenon. They would not agree about a hypothetical water bucket or a lonely planet in an otherwise empty universe, but not because of conflicting intuitions, but because Mach would rather refuse to trust his unreliable intuitions about the case. In the counterfactual case, it could indeed come to a clash of intuitions. It seems that Mach was aware of this and *therefore* refused to accept the thought experiment of Carl Neumann. Let us now turn to Einstein.

Einstein seems to have a similar situation in mind. Two fluid bodies, S1 and S2, of the same size are hovering freely in space, rotating with constant angular velocity about the line joining the masses (as judged by an observer at rest relatively to the other spheroid, respectively). Each spheroid was measured and S1 proved to be a sphere, S2 proved to be an ellipsoid of revolution. Einstein argues that the difference in shape *can* and *should* be explained without postulating absolute space, but by postulating the existence of Galilean systems, which are *observable*.

This seems to be essentially the same argumentative situation as it was with Mach. The difference between Einstein and Newton is not a difference in intuitions concerning the details of the case imagined, but the question what would serve as the best explanation for such a phenomenon. This is in no way a special problem of thought experimentation, but the old problem of how to decide between empirically equivalent competing hypotheses. It is also clear that Einstein does not present a thought experiment to prove or refute a theory here, he wants to "illustrate" by way of example a feature of classical mechanics and special relativity theory.[16]

II.2 *EPR, Bell Inequality and Efficient Detection*

[15] But see Sorensen 1992 for a discussion.
[16] Einstein 1920, 82.

When Albert Einstein, Boris Podolsky and Nathan Rosen set up their thought experiment, they did not try to refute quantum mechanics, but to criticize a certain interpretation of it, mainly the Copenhagen view defended by Niels Bohr, according to which quantum mechanics can be considered a *complete* theory of physical reality. The thought experiment was thus supposed to uncover an *imperfection* of quantum mechanics.

What EPR tried to show was that given a certain criterion of physical reality, QM allows to deduce the existence of more elements of physical reality than QM seems to account for. The criterion (meant as a sufficient condition only) was this:

> If, without in any way disturbing a 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/Rosen 1935/1983, 138.)

The argument proceeded by postulating a certain experimental setup and showing that QM would predict for this setup with certainty the value of two physical quantities P and Q, although QM claims the same time that not both quantities have simultaneous physical reality:

> One could object to this conclusion on the grounds that our criterion of reality is not sufficiently restrictive. Indeed, one would not arrive at our conclusion if one insisted that two or more physical quantities can be regarded as simultaneous elements of reality *only when they can be simultaneously measured or predicted*. On this point of view, since either or the other, but not both simultaneously, of the quantities P and Q depend upon the process of measurement carried out on the first system, which does not disturb the second system in any way. No reasonable definition of reality of P and Q could be expected to permit this. (Einstein/Podolsky/Rosen 1935/1983, 141)

What Bohr did in reaction to the thought experiment was to try to find a flaw in the argument by Einstein, Podolsky, and Rosen. He didn't thereby invoke an alternative set of intuitions about what happens in the experiment, but tried to point out an argumentative mistake in the argument, he tried to show that Einstein, Podolsky, and Rosen fell prey to an ambiguity in their notion of 'criterion of reality':

> In fact, as we shall see, a criterion of reality like that proposed by [Einstein, Podolsky, and Rosen] contains – however cautious its formulation may appear – an essential ambiguity when it is applied to the actual problems with which we are here concerned. (Bohr 1935, /1983, 145)

Whether Bohr succeeded in pointing out a flaw in the argument by EPR is highly questionable. The success of his argument might best be explained sociologically.[17] But that is not the point. Bohr did not have contradicting "intuitions" about what would be happening in the case and his reply wasn't based on them. He tried to *argue* against the EPR conclusion, by pointing out what he thought is an ambiguity in the original argument.

The conflict is in fact not about what *happens* in the imaginary experiment. EPR and Bohr are in agreement about what all measurements will be like if such an experiment could be carried out. The conflict is about what counts as a *satisfying* explanation of such a phenomenon. Einstein is again invoking an epistemological principle, which led him to favour a hidden-variables theory. This alternative hypothesis is taken to be empirically indistinguishable from the Copenhagen interpretation of QM. That *this* assumption is false was the insight of Bell, 30 years later.

So far we have discussed the question of whether or not there are scientific thought experiments that trigger contradicting intuitions in the audience and whether or not these arguments were question begging. We found in the examples referred to by PA no support for any of these claims. We shall now turn to the remaining question of whether an *experimentum crucis* in physics could settle the dispute over a physical thought experiment, (PA5).

As PA implicitly admit themselves, this is not the case for all physical thought experiments. It seems clear, e.g., for Galileo's Pisa-experiment that an empirical test could not have solved the matter the way his thought experiment did.[18] But this point is not very remarkable, given that PA argue elsewhere that Galileo's Pisa-experiment was a poor one in the first place (Atkinson/Peijnenburg forthcoming). Therefore we shall turn to EPR and the "crucial experiment" carried out by Aspect in 1982. PA suggest that this crucial experiment brought the conflict in physics to a halt, in a way that could not be expected to happen in philosophy. The situation in philosophy would continue but shift to being about the correct interpretation of the experiment.

First of all, Aspect didn't immediately bring EPR to a test. EPR – as PA admit – was first "retooled" by David Bohm in the 50ies. This made way for John Bell to formulate his famous inequalities in 1964, which in fact lead to a predictable empirical difference between certain hidden variables theory and the Copenhagen view. Before that, EPR could not be brought to a test, for it was unclear what the *empirical difference* of both views would amount

[17] See also Düsberg 1998.
[18] But see Kühne [manuscript] for a well argued critique of this statement.

to. From that point it took another couple of years until Aspect found a way to test the Bell inequalities in the early 80ies.

The second important point to note about the recent history of physics is that this did not close the topic. Since then the Aspect experiments are under discussion, for there are at least three loopholes for a "local realist"[19]:

(i.) the locality or "lightcone" loophole

The correlations which support the Copenhagen interpretation could result from unknown subluminal signals propagating between different regions of the experiment apparatus. (Thus the way the apparatus is set up still allows for speed of light, or slower than speed of light communication.)[20]

(ii.) the detection loophole

Experiments test the Bell inequalities by assuming that they measure a fair sample of the particles. However, most experiments have detection efficiencies low enough to allow for the possibility that the ensemble of detected events actually measured agrees with the Copenhagen interpretation, whereas the entire ensemble (all events together) satisfy the Bell inequalities.[21]

(iii.) the "accidentals" loophole

This objection is not to Bell tests in particular, but to the experimental procedure adopted in it. After the measurement, so called "accidentals" are subtracted from the empirical findings (to account for artefacts of the measuring process). After the subtraction, the coincidences are analyzed. Some data suggests that without the subtraction the results would not have violated the Bell inequalities.[22]

In a recent physics paper on such loopholes, the author comes to the following conclusion:

> So why have EPR experiments been so widely accepted as supporting QT, when there are perfectly straightforward local realist possibilities, just a few of which I have introduced in this paper? […]

[19] Whether all or any of them are or were *real* loopholes is or was, of course, subject of discussion. My point is not that after Aspect's experiments *was* room for doubt in the sense that there were objective epistemic possibilities open, but there were *subjective* possibilities left open for some physicists which is clear from the discussion.
[20] Weihs et al. 1998, 5039.
[21] Rowe et al. 2001, 791.
[22] Thompson [manuscript].

> The explanation for this phenomenon lies in sociological and psychological factors – confusion caused by working with a counterintuitive theory, the pressure to produce results acceptable to peers, the conviction that nobody else has yet found fault with QT. One "success" in EPR experiments has led to another, but the faults have propagated instead of weeded out. (Thompson [manuscript], 4)

Do not get me wrong, I'm not endorsing any such view. What I tried to show in this section is that the situation in physics is not so much unlike philosophy. Even if – after a period of non-trivial reformulations – a thought experiment can be tested empirically, this doesn't necessarily close the discussion in physics. As in philosophy the discussion might continue and shift to a discussion about the correct interpretation of the experiment. I would predict that the likelihood for this is much higher, if the real experiment refuted the conclusion drawn from the thought experiment.[23]

PA might reply this: 'OK, so the difference between physics and philosophy is merely quantitative, not qualitative, in this respect. That doesn't really matter. The difference between a molehill in my backyard and the Mount Everest is "merely" quantitative too. The important point is that in physics most scientists can still *by and large* agree about the interpretation of an experiment or the merits of a background theory, whereas in philosophy there is *no* way towards a consensus.'

Moreover, my discussion of the examples of physics have so far only established that *in physics* does a clash of intuitions not seem to occur. It might still be that the situation in philosophy is worse. Thus in the next section I will show that the difference between philosophy and physics when they deal with thought experiments is misconceived, by showing how philosophers are trying and succeeding to reach agreement over the issues involved.

III. The Situation in Philosophy: Why Zombies are not a Matter of Taste

Now let's see in some more detail what is going on in the philosophical examples given. For simplicity we will concentrate on Frank Jackson's Mary and David Chalmers' zombie. We have to begin with a brief recapitulation of what is at issue, respectively:

III.1 *Mary and what it is like to see red*

Mary, the colour scientist was raised in a black and white room. She learned everything there is to know about the neurophysiological account of what it is like to see red. (Mary is living in

[23] I'm assuming that in science no contradicting conclusions are drawn from a thought experiment, thus my prediction here is not trivial. I argued for this assumption above.

a future in which neurophysiology is "complete".) She knows what happens in the brain and elsewhere when people see red things, in particular what the specific brain state is people are in when they see red things. Now, for the first time she is allowed to leave her prison and sees a ripe tomato. She forms a new thought: 'This is what it is like to see red.'

The intuition triggered is that Mary learned something new that she couldn't know from the neurophysiology textbooks she had read in her prison. Although she is an expert about colour experiences, there was something not accounted for in her textbooks that she now learned. Thus there seem to be facts about colour vision that cannot be explained by physicalism. Physicalism claims that everything there is to know about the world (including everything there is to know about experiences) is entailed in the complete physical account of the world. Mary had access to the complete physical account of experiencing red things, but she nevertheless learned something new, thus physicalism is false.

Note, that the intuition triggered is that Mary learned something new! The intuition is not that physicalism is false. This only follows after adding *other* assumptions. It is correct that some philosophers have challenged the thought experiment on the grounds that it is impossible to arrive at a verdict about whether Mary learned something new, because it is unclear what it means for neurophysiology to be 'complete'. Since we don't know what a completed neurophysiology might look like, we must rely on abstract characterizations, but there is a problem: if 'complete' means 'everything there is to know about colour experiences', Mary cannot learn anything new by definition. If it means 'everything neurophysiology can tell you, i.e. everything there is to know except this kind of new stuff Mary learns' the thought experiment is question begging. This line of thought does not criticize the thought experiment by relying on contradictory intuitions, however! The worry is that the thought experiment cannot show what it is supposed to show for its description is question begging.

Other philosophers have challenged the move from 'Mary learns something new about colour experiences' to 'There is a fact about colour experiences that was left out by the neurophysiological description'. These philosophers try to analyze Mary's new knowledge as *knowing how* rather than *knowing that*[24], or analyze Mary's knowledge as concerning a new aspect of a fact that she knew about already in neurophysiological terms[25]. Again, these moves do not rely on different intuitions about the case! The intuition ('Mary learned something new.') is shared, what is doubted is what conclusions are to be drawn from it. This discussion proceeds by comparing different theories of knowing how and knowing that, as well as different theories of the content of propositional attitudes and comparing them in the

[24] See Perry 2001 on Lewis/Nemirow (152-159).
[25] Perry 2001, 166.

light of independent evidence. Thereby philosophers might again use intuitions in their arguments but these intuitions are neither about Mary, nor do they contradict each other. Therefore, in the Mary case (PA1) and (PA4) do not apply.

(PA2), which was the thesis that thought experiments in philosophy are question begging, might apply in a way, for it is considered to be a possible flaw of the thought experiment *by participants of the debate*. Thus, if (PA2) should apply, this is considered a flaw by the methodological standards of the profession, thus not a general methodological problem with thought experimentation in philosophy (it were, if the profession had no such internal regulations in their methodological practice). The question of whether or not the Mary thought experiment is question begging does not depend on any intuitions, but on whether or not the notion of 'completeness' can be explicated independently. This is way different from PA's analysis of the case. There is no reason to assume that the discussion should be doomed to go "continually round on a merry-ground".

III.2 *Zombies and what they tell us about physicalism*

The situation with zombies is similar. Modern physicalists do in general not subscribe to the view held by members of the Vienna Circle that every statement about a mental state is analytically transformable into a statement about physical states. Modern physicalists are in general "a posteriori physicalists"[26], i.e. they believe that we found *a posteriori* mental properties to supervene strongly on physical properties (or that mental states are a posteriori identical with physical states). If we *had* to find things *a posteriori* to be in a certain way, it must at least be logically possible that things could have been otherwise. Therefore it is generally agreed to be a logical possibility that zombies exist. It is conceivable that an exact physical duplicate of our world could lack all consciousness. The question is what follows from this shared intuition.

Some hold that logical possibilities cannot cut any metaphysical ice. Some think they can. Chalmers and Jackson have further developed a semantico-epistemological theory that is supposed to explain how conceivings of certain kinds can reveal what is logically possible and how what is logically possible is connected to what is metaphysically possible. This is so called two-dimensionalism.[27] If two-dimensionalism (as presented by Chalmers) is the most plausible account of logical and metaphysical possibilities, we should conclude that zombies

[26] This is an oversimplification of the issue. It is sufficient to make my point here. Getting more into details would only prove that the philosophical landscape and the arguments used are of a richer variety than it is suspected by Atkinson and Peijnenburg.
[27] See Chalmers 2002, 2004, Jackson 1998, and Cohnitz 2003a for a discussion.

are metaphysically possible. But physicalism was supposed to be a metaphysical thesis that holds with metaphysical necessity. If zombies are metaphysically possible, physicalism must be false.

Note, two-dimensionalism is not about physicalism or neo-dualism but a modal epistemology that makes certain assumption about semantics! When it is discussed whether zombies can prove that physicalism is false, what is discussed is whether a transition from logical to metaphysical possibilities can be justified along the lines Chalmers and Jackson suggest. Again, the discussion in philosophy is not a mere duel of intuitions, pro-zombie on the side of neo-dualists, anti-zombie in the physicalist camp. Philosophers are discussing whether the background theories that explain the intuitions (or explain the intuitions away) are plausible. These discussions are independent of the target issues (which is physicalism in the examples considered). (PA1) and (PA4) do not apply.

To make sure, some philosophers have argued, that the zombie thought experiment is question begging. Again, this is an *accusation*. It is not the case that a physicalist would use the zombie thought experiment to argue for physicalism by claiming he had just the opposing intuitions. This is a distorted picture Peijnenburg and Atkinson are suggesting.

In fact, the argument why the zombie thought experiment is question begging is similarly subtle as in the Mary case. In the zombie thought experiment we are asked to consider an exact physical duplicate of our world. Let's first assume mental causation, i.e. that mental states can cause physical states. Say the mental state (M) I was in when I by mistake touched the hot plate yesterday caused the physical event (P) that immediately followed, i.e. the moving of my hand from the hot plate. In a physical duplicate world (w_2) of our world (w_1) without any consciousness, (M) is not present. But then what about (P)? (P) should be in (w_2) – it is a physical event – , although uncaused by (M). How can that be? Did (P) just occur uncaused at all? That shouldn't be the case, for (w_2) was supposed to be a physical duplicate world and in our world (w_1) physical events don't just occur. Thus how can (P) be part of a physical duplicate world (w_2), if that world lacks (M)?

It seems we have to give up the assumption of mental causation about our world (w_1) to coherently assume that (P) can be part of (w_2) although (M) is no part of it. The description of the thought experiment seems to presuppose some kind of epiphenomenalism. If physicalism implies the negation of epiphenomenalism, the description of the thought experiment might beg the question against the physicalist.[28]

[28] See Perry 2001. It is questionable whether Perry really can show that his objection undermines the argument. See also Chalmers forthcoming.

Again, if the thought experiment is question begging, then this is noticed within the profession. Thus, the professional discussion in philosophy seems to be sufficiently equipped with *internal* correcting mechanisms that will keep it in general from the disastrous consequences Peijnenburg and Atkinson foresee.

IV. Conclusion

As we have seen, philosophy is not a mere battle of intuitions. It nevertheless often appears to be, because bats, zombies, teletransporters, and brain transplants are more impressive than details about two-dimensional modal logic or theories of propositional attitudes. But these are the background-theories we turn to and discuss when a thought experiment threatens to refute our cherished theory. They are usually independent of the theory under attack, thus philosophy is not in general question begging. If it is, this is detected by philosophers and used as an argument (this couldn't work, if it were general practice to use question begging thought experiments). Lay persons might not take notice of this and be blinded by the outrageous stories philosophers seem to contemplate all day.

PA have left us with indicators for "poor" thought experiments. But when is a thought experiment a good one? The idea of Atkinson was that a thought experiment is a good one if it can be turned into a real experiment. What to say about this?

First of all, whether a thought experiment is a good one then turns out to be a question we can never answer in advance.[29] When they designed the EPR thought experiment, Einstein, Podolsky, and Rosen couldn't foresee Bohm's reformulation nor Bell's inequalities. EPR turned out to be "good" 50 years after it was designed. That is pretty unhelpful for methodological considerations. This is also basically the reason why I did not discuss PA's unhelpful suggestions how philosophy should "improve". If we knew how to settle philosophical problems in every case empirically, we would do so. Fact is that for most of our problems we have absolutely no clue in what way they might connect to empirical findings.

Moreover, Atkinson judges the EPR thought experiment "good" independent of its content and target. EPR was designed to *refute* the Copenhagen interpretation, but the empirical test rather *corroborated* it. But if our criterion for a "good" thought experiment is indifferent with respect to the content and target of a thought experiment, then thought

[29] Note that in Atkinson's and Peijnenburg's analysis the two indicators are not conclusive. A thought experiment with both features, e.g. EPR, which Atkinson and Peijnenburg assume to have "contradicting" as well as "question begging" conclusions, can nevertheless be a "very good one" if it can later be turned into a real experiment. Thus, whether a thought experiment is a bad one turns out to be something that we *never* know (it might always be that we can turn the inconclusive thought experiment into a conclusive real experiment next week).

experimentation is methodologically speaking on a par with sports, sex, and good coffee, which all might equally be good things to do or to have if they lead to fruitful experimentation (50 years later, maybe). They are mere practices and the hope that we can say anything methodologically enlightening about them is dim. So what are good thought experiments, really? I am happy to tell you elsewhere.

V. References

Atkinson, David, 2003, Experiments and Thought Experiments in Natural Science, M.C. Galavotti (ed.), Observation and Experiment in the Natural and Social Sciences, Boston Studies in the Philosophy of Science, Vol. 232, Dordrecht: Kluwer, 2003, 209-225.

Atkinson, David/Peijnenburg, Jeanne, [forthcoming], *Galileo and Prior Philosophy*. To be published in *Studies in History and Philosophy of Science*. http://www-th.phys.rug.nl/~atkinson/galileo.pdf

Bealer, George, 1998, Intuition and the Autonomy of Philosophy, Michael R. DePaul and William Ramsey (eds.), *Rethinking Intuition, The Psychology of Intuition and Its Role in Philosophical Inquiry*, 201-239.

Bohr, Niels, Can quantum-mechanical description of physical reality be considered complete? J. A. Wheeler, W. H. Zurek (eds.), *Quantum Theory and Measurement*, Princeton University Press 1983, 145-151.

Brendel, Elke, 1999, Gedankenexperimente als Motor der Wissenschaftsdynamik, J. Mittelstraß (ed.), *Die Zukunft des Wissens* (XVIII. Deutscher Kongress für Philosophie Konstanz 1999), Konstanz, Universitätsverlag Konstanz.

Chalmers, David, 1996, *The Conscious Mind: In Search of a Fundamental Theory*, OUP 1996.

Chalmers, David, 2002, Does Conceivability Entail Possibility? Tamar Szabó Gendler and John Hawthorne (eds.), *Conceivability and Possibility*, Oxford, Clarendon Press.

Chalmers, David, [forthcoming], The Foundations of Two-Dimensional Semantics, to appear in M. Garcia-Caprintero and J. Macia, (eds.), *Two-Dimensional Semantics: Foundations and Applications*, Oxford Univesity Press, 2004. http://www.u.arizona.edu/~chalmers/papers/foundations.html.

Chalmers, David, 2002, On Sense and Intension, *Philosophical Perspectives* 16, 135-182.

Chalmers, David, [forthcoming], Imagination, Indexicality, and Intensions. To be published in *Philosophy and Phenomenological Research*. http://www.u.arizona.edu/~chalmers/papers/perry.html

Cohnitz, Daniel, 2003a, Two-Dimensionalism and the Metaphysical Possibility of Zombies, B. Löwe, W. Malzkorn, T. Räsch (eds.), *Foundations of The Formal Sciences II*. Applications of Mathematical Logic in Philosophy and Linguistics [Trends in Logic]. Dordrecht, Kluwer Academic Publishers.

Cohnitz, Daniel, 2003b, Modal Skepticism: Philosophical Thought Experiments and Modal Epistemology, *The Vienna Circle and Logical Positivism* [Vienna Circle Institute Yearbook 10/2002]. Dordrecht: Kluwer Academic Publishers 2003, 281-296.

Cohnitz, Daniel, 2003c, Personal Identity and the Methodology of Imaginary Cases, Klaus Petrus (ed.): *On Human Persons*. Ontos Verlag 2003, 145-181.

Düsberg, Klaus-Jürgen, Probleme einer Interpretation der Quantenmechanik. Zur Relevanz der Quantenmechanik für die Philosophie, Jochen Lechner (ed.), *Analyse, Rekosntruktion, Kritik. Logisch-philosophische Abhandlungen* [Studia Philosophica et Historica 23], Peter Lang Verlag 1998, 243-288.

Einstein, Albert, 1920, Die Grundlage der allgemeinen Relativitätstheorie, H.A. Lorentz et al. (eds.), Das Relativitätsprinzip. Eine Sammlung von Abhandlungen [Fortschritte der mathematischen Wissenschaften in Monographien 2], Teubner-Verlag 1920, 81-124.

Einstein, Albert/Podolsky, Boris/Rosen, Nathan, 1935/1983, Can quantum-mechanical description of physical reality be considered complete?, J. A. Wheeler, W. H. Zurek (eds.), *Quantum Theory and Measurement*, Princeton University Press 1983, 138-141.

Gähde, Ulrich, 2000, Gedankenexperimente in Erkenntnistheorie und Physik: strukturelle Parallelen, Julian Nida-Rümelin (ed.), *Rationalität, Realismus, Revision : Vorträge des 3. internationalen Kongresses der Gesellschaft für Analytische Philosophie vom 15. bis zum 18. September 1997 in München*, Berlin, de Gruyter, 457-464.

Jackson, Frank, 1998, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford, Clarendon Press.

Kühne, Ulrich, [manuskript], *Die Methode des Gedankenexperiments: Untersuchung zur Rationalität naturwissenschaftlicher Theoriereformen*, Dissertation, Universität Bremen.

Leibniz, G. W., 1686/1996, *Philosophische Schriften*, Vol. 1, edited and translated by Hans Heinz Holz, Frankfurt a.M., Suhrkamp.

Locke, John, 1975, Of Identity and Diversity, John Perry (ed.), *Personal Identity*, Berkeley, University of California Press, 33-52.

Mach, Ernst, 1883/1988, *Die Mechanik in ihrer Entwicklung. Historisch-kritisch dargestellt*. Akademie-Verlag 1988.

Neumann, Carl, 1870, *Die Prinzipien der Galilei-Newtonschen Theorie*, Teubner-Verlag 1870.

Newton, Isaac, 1726, Philosophiae naturalis principia mathematica. Editio tertia aucta & emendata. Londini: Apud Guil. & Joh. Innys, Reggiae Societatis typographos 1726.

Noonan, Harold, 1991, *Personal Identity*, Routledge 1991.

Peijnenburg, Jeanne/Atkinson, David, 2003, When are thought experiments poor ones? *Journal for General Philosophy of Science* 34, 2003, 305-322.

Perry, John, 2001, *Knowledge, Possibility, and Consciousness*, MIT Press 2001.

Perry, John, 2002, *Identity, Personal Identity, and the Self*, Indianapolis/Cambridge, Hackett.

Putnam, Hilary, 1990, *Die Bedeutung von "Bedeutung"*, Klostermann 1990.

Rowe, M.A. et al., 2001, Experimental violation of Bell's inequality with sufficient detection, *Nature* 409, 791-794.

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

Swinburne, Richard G., 1974, Personal Identity, *Proceedings of the Aristotelian Society* 74, 1973-74, 231-247.

Thompson, Caroline, Timing, [manuscript], *"Accidentals" and Other Artifacts in EPR Experiments*. http://xxx.lanl.gov/PS_cache/quant-ph/pdf/9711/9711044.pdf

Torretti, Roberto, 1999, The Philosophy of Physics, Cambridge University Press 1999.

Weihs, Gregor, et al., 1998, Violation of Bell's inequality under Strict Einstein Locality Conditions, Physical Review Letters 81, 1998, 5039-5043.

Yablo, Stephen, 1993, Is Conceivability a Guide to Possibility, *Philosophy and Phenomenological Research* LIII, 1-42.