Intuition Pumps and the Proper Use of Thought Experiments

Elke BRENDEL[†]

ABSTRACT

I begin with an explication of "thought experiment". I then clarify the role that intuitions play in thought experiments by addressing two important issues: (1) *the informativeness of thought experiments* and (2) *the legitimacy of the method of thought experiments* in philosophy and the natural sciences. I defend a naturalistic account of intuitions that provides a plausible explanation of the informativeness of thought experiments, which, in turn, allows thought experiments to be reconstructed as arguments. I also specify criteria for distinguishing bad "intuition pumps" from legitimate thought experiments. These criteria help us to avoid being seduced by the dangerous suggestive power of misleading intuitions.

1. Introduction

Although thought experiments are completely carried out in the "laboratory of the mind", they are important methodological instruments for scientific inquiries and can give us new insights about the world. In order to do so, thought experiments must appeal to *intuitions*. But what is the exact role that intuitions play in thought experiments?

There are mainly two important problems in the philosophical discussions of thought experiments where intuitions are involved: First, there is the problem of the informativeness of thought experiments. Thought experiments provide us with new information. But where does such information come from? The problem of the informativeness of thought experiments is most controversially discussed in the debate between James Robert Brown and John D. Norton. Brown supports a Platonic account of thought experiments in order to solve the informativeness problem. He holds that in some thought experiments we gain new information through the help of intuitions which are irreducible to and independent of empirical, inferential or any other *a posteriori* knowledge. The intuitions that emerge in these thought experiments give us, according to Brown, *a priori* access to a metaphysical realm of universals and

[†] Universität Mainz. Email: brendel@uni-mainz.de

the relations between them. In sharp contrast to this position, Norton argues that thought experiments are indeed just (deductive or inductive) arguments. We therefore gain information in thought experiments quite naturalistically through sense experience and inference. Second, there is the problem of misusing thought experiments as intuition pumps. This notion comes from Daniel Dennett who stresses that the highly imaginative scenarios of some thought experiments can distract from a thorough examination and critical reflection of thought experiments. By appealing to our intuitions, thought experiments can lead us to a quick and uncritical jump to a conclusion that is not really warranted.

In the following I will address both of these problems. First of all, I will offer an explanation of "thought experiment" which to my judgement provides the most philosophically fruitful understanding of this concept. After that I will turn to the problem of informativeness and argue for a specific *naturalistic* account of intuitions. This account gives a plausible explanation for some obvious characteristics of intuitions, like their fallibility, relative instability and fragility in some areas of scientific inquiry. It can also explain the informativeness of some thought experiments without referring to a distinctively non-empirical philosophical method of intellectual insight and without postulating a Platonic ontology of universals and objective laws of nature. With Norton I will hold that thought experiments can be reconstructed as mere arguments.

The main part of my paper will be devoted to the problem of the legitimacy of the method of thought experiments. In particular, I will formulate some criteria which can help us separate bad intuition pumps from those thought experiments that are legitimate and rationally justified arguments and which therefore can help us avoid the trap of the dangerous suggestive power of intuitions.

2. What are thought experiments?

In a paper dated 1811 Hans Christian Ørsted was the first person who mentioned the term "thought experiment" as a separate source of knowledge.[1] But it was Ernst Mach who coined the term "thought experiment" for the philosophical discussion.[2] Mach uses the term in a very wide sense. According to him, thought experiments can be almost all kinds of "thought experiences",

[1] See Ørsted 1811.
[2] See Mach 1897.

like dreaming, hallucinating, writing novels or imagining utopia.[3] Thought experiments also have the important propaedeutic function of being a necessary precondition for planning and executing real experiments. But this very general conception of thought experiments does not adequately reflect the fact that a thought experiment is indeed a certain kind of *experiment.* Although it is an imaginary investigation that need not or cannot be executed in the real physical world, it is nevertheless subject to certain theoretical requirements that it shares with real experiments. For example, a "thought experimenter" also studies the functional dependency of variables by planned and controlled data change. Furthermore, in a manner similar to real experiments, every thought experiment depends on some background assumptions or background theories.

The main difference between thought experiments and real experiments lies in the fact that according to the intentions of the "thought experimenter", the aims of the thought experiment can be achieved without needing to perform a real experiment. In contrast to a real experiment (or to a simulation), the supposed outcome of a thought experiment is not open but can be "seen" by some intuitional insight. Some thought experiments cannot be realised since they make use of unreal situations or depend on counterfactual, fictional or idealized assumptions – like Maxwell's demon, Einstein's observer travelling besides a beam of light, Parfit's teletransporter or Galileo's bodies falling without any air resistance. But there are also thought experiments that could be executed by a real physical experiment (for example, Newton's bucket or Einstein's train experiment) or in which a real or possible situation is imagined. A lot of thought experiments in ethics are of the latter kind. Therefore, counterfactuality is not a necessary condition of a thought experiment.[4] But what are the primary aims and purposes of a thought experiment? What is its scientific function?

According to Sorensen, thought experiments are to be regarded as "alethic refuters" whose main function consists in the proof of paradoxes: "Picture thought experiments as expeditions to possible worlds. The mission is to refute a source statement that has an implication about the constituents of these worlds." (Sorensen 1992, 135) Thomas Kuhn also stresses this destructive element of thought experiments. With the help of thought experiments, we can

[3] See Mach 1905, 183.

[4] My usage of the term "counterfactuality" is restricted to situations that are *physically impossible*, i.e., situations that contradict the laws of nature. In a "counterfactual assumption" it is thus assumed that a physically impossible situation holds. A mere physically possible situation that is not (or not yet) actual is – according to this strict sense of "counterfactuality" – *not* counterfactual.

track down hidden contradictions. In particular, these contradictions arise when we imagine new or unusual situations in which the criteria that govern the use of concepts in familiar areas break down and lead to paradoxical results.[5] The uncovering of those inconsistencies is important for the rational progress of scientific theories, and it is surely one of the main functions of thought experiments. But there are also other purposes of thought experiments. Thought experiments can be used to provide evidence in support of a questionable theory. Newton's bucket experiment, for example, was intended to show the existence of absolute space. Furthermore, thought experiments can have the pedagogical function of illustrating an otherwise complex and abstract position. John Locke's "prince and cobbler" thought experiment, for example, was intended to illustrate his view that psychological continuity is a necessary condition for personal identity since the identity of a person means "sameness of a rational being".[6] Other thought experiments detect vagueness or the borderline cases of concepts, as for example those thought experiments that show that we don't have stable intuitions concerning the concept of the identity of objects or persons or the concept of life. By this means, a thought experiment can help to explicate a concept and its area of application more precisely.

To sum up the explication of the concept "thought experiment" I have provided so far:

- Thought experiments are intended to achieve their aims without the need of their execution in a real physical experiment.
- Thought experiments are *experiments* insofar as they share certain minimal theoretical requirements with real experiments, like
  a) the planned and controlled change of data,
  b) showing in an artificial situation how variables are functionally dependent on each other, and
  c) their dependency on some background hypotheses or background theories in order to analyse and evaluate the experiment.
- Thought experiments have the following functions and purposes:
  a) to prove that certain theories or concepts involve contradictions,
  b) to give supporting evidence for a theory or a concept,
  c) to illustrate a complex or abstract position, and/or
  d) to detect vagueness or the borderline cases of a concept.

Given this explication of "thought experiment", we can now turn to the problem of informativeness.

[5] See Kuhn 1977.
[6] See Locke 1690, book II, ch. 27.

3. The problem of the informativeness of thought experiments

How can thought experiments fulfil the functions mentioned above? How can we learn something new about theories and concepts, if a thought experiment clearly involves no empirical input? Where does the information provided by the thought experiment come from? In his bold answer to these questions James Robert Brown assumes that a certain class of thought experiments appeals to a special intuitive faculty which functions as a vehicle for direct access to universals and laws of nature. This certain class consists of so-called "Platonic" thought experiments which Brown describes as follows:

> A *Platonic thought experiment* is a single thought experiment which destroys an old or existing theory and simultaneously generates a new one; it is *a priori* in that it is not based on new empirical evidence nor is it merely logically derived from old data; and it is an advance in that the resulting theory is better than the predecessor theory. (Brown 1991a, 77)

According to Brown, there are various examples of such astonishing thought experiments, but "the greatest example of all" (Brown 1991b, 125) is Galileo's famous thought experiment to show that regardless of their weights all bodies fall at the same speed. In the *Discorsi* Galileo first states Aristotle's view that different bodies in a given medium fall in proportion to their weights with different speed. For example, a body A which is 10 times heavier than a body B falls ten times faster than B.[7] But this assumption leads to an absurd consequence: If we combine a heavier body A with a lighter body B, B will decelerate A, and therefore, the combined system A + B will fall more slowly than A alone. But on the other hand, A + B is heavier than A alone and should consequently fall faster than A alone. So we have the inconsistent result that A + B falls both faster and slower than A alone. Therefore, the Aristotelian view of falling bodies has been destroyed, and, as Brown remarks, it is now "plain as day" that we have to resolve the paradox by embracing the new theory in which all bodies fall (in a given medium) at the same speed.[8]

According to Brown, this wonderful thought experiment is destructive and constructive at the same time: It destroys the old Aristotelian theory and simultaneously jumps immediately to the new theory. This immediate jump is, as Brown puts it, "a case of *a priori* science" (Brown 1991b, 125). The step from the contradiction to the conclusion that all bodies fall at the same speed cannot, according to Brown, be considered as any kind of argumentation since

[7] Galileo 1638, 334.

[8] See Brown 1991b, 123.

the transition from the old to the new theory is driven neither by empirical input nor by logical deduction from old data, nor by "making the simplest overall adjustment to the old theory" (Brown 1991b, 125). Instead, we obtain *a priori* knowledge by intuitive perception of the relevant laws of nature (in this given example, by the laws that govern falling bodies). Brown obviously supports Platonism and a realist account of laws of nature. He believes that there is a special faculty of intuition that reveals itself in Platonic thought experiments. Therefore, Platonic thought experiments provide us with a distinctive method – different from the empirical or inferential methods of scientific enterprises. Nevertheless, Brown's Platonism allows for the fallibility of *a priori* knowledge. The intuition that emerges in Platonic thought experiments can at times be misleading and unreliable as a method of gaining access to the Platonic world. He writes:

> Unlike Plato, Descartes, or Leibniz, etc., *a priori* knowledge on my view is neither certain nor innate. It is not put there by God; it is not remembered; nor is it infallible. But like the traditional rationalists, I hold that the abstract realm is perfectly real and that we can know something about it. (Brown 1991b, 127)

I find Brown's Platonic account of thought experiments highly implausible. I will respond in agreement with Norton's claim that thought experiments can always be reconstructed as arguments and that there is no immediate grasp of the relevant laws of nature in the realm of universals, i.e., there is no conclusion of a thought experiment undermined by the premises of the argument.[9]

As far as Galileo's thought experiment is concerned, the supposedly intuitive grasp of the conclusion that all bodies fall at the same speed does not at all follow immediately and with no help of other premises from the demonstrated contradiction. The consequence is, for example, only conclusive, if we make the further assumption that we can ignore the other components of the falling bodies – like their shape, their material etc. – and that it is only the weight that is responsible for the speed of the falling bodies. But this additional assumption is for arbitrary media false, as Salviati in the *Discorsi* had already remarked.[10] That all bodies fall at the same speed if one completely neutralizes the air resistance was nothing more than a conjecture in Galileo's day. The controversial dispute about the interpretation of this thought experiment in the *Discorsi* shows that no one had an immediate grasp of the real law of falling bodies involved. When we in the modern era – and Brown in par-

[9] Cf. Norton's "reconstruction thesis", in Norton 1996, 339.

[10] See Galileo 1638, 65.

ticular – look at this thought experiment from an historically distant perspective and with the knowledge of modern physics concerning falling bodies in a vacuum, the inference from the contradiction to the "right" conclusion just *seems* to be immediate and untutored by any empirical or logical reasoning, since we implicitly assume that the experiment must be executed in a medium where we can ignore air resistance. So the intuition that leads us to the conclusion that all bodies fall at the same speed is dependent on and controlled by our empirical and scientific knowledge.

I also find Brown's fallible Platonism unintelligible because it remains entirely unclear when and why an intuitive grasp of the abstract realm can go wrong. On the other hand, the question of whether a thought experiment is legitimate or not can be answered much more adequately if we treat the thought experiment as a reconstructed argument. The justification of a thought experiment depends on this account on the justification of the premises involved and the conclusiveness of the inferential steps.

These considerations show first of all that Brown's favourite example of a Platonic thought experiment can be reconstructed as a *reductio ad absurdum* argument by uncovering its hidden premises.[11] That is why there is really no argumentative gap in the thought experiment needing to be surmounted by an intuitive leap. Second, we have shown that this intuitive leap cannot be construed as a specific *a priori* method of gaining insight into the realm of a Platonic world, a world in which the laws of nature somehow exist independently of us. Instead, the intuitions that lead to the deduction that all bodies fall at the same speed have to be justified in an *a posteriori* way by appealing to our former knowledge about falling bodies. Therefore, we should reject the view that we gain new information from thought experiments by a special epistemic capacity for intuitively perceiving the laws of nature in a Platonic realm. Instead, in thought experiments we gain new information by rearranging or reorganizing already known empirical data in a new way and drawing new inferences from them or by looking at these data from a different and unusual perspective. In Galileo's thought experiment, for example, the rearrangement of empirical experience consists in the original idea of combining bodies of different weight. In order to explain the informativeness of thought experiments, no mysterious access to the Platonic realm needs to be postulated. The way that we get new information through a thought experiment can be modelled entirely as a logical argument.

[11] For details of reconstructing this thought experiment as an argument see Norton 1996.

Although thought experiments can be *reconstructed* as arguments, I do not want to go as far as Norton and claim that all thought experiments *really are explicit* arguments, i.e. that "the actual conduct of a thought experiment consists of the execution of an argument" (Norton 1996, 354). If there is in fact always a complex kind of argumentation going on in our minds while we conduct a thought experiment, the effortless, quick and very often commonly shared conclusions drawn from thought experiments would be difficult to explain. I think it is more plausible to assume that some hidden premises function as background knowledge, and that they need not be explicitly activated when we conduct a thought experiment.

Part of this background knowledge consists of intuitions. Intuitions are to my mind best regarded as mental propositional attitudes which are accompanied by a strong feeling of certainty. In thought experiments they can be part of the background knowledge and therefore may at times unconsciously determine our interpretation. But, as said above, they do not provide a special method of *a priori* access to the Platonic realm. Some intuitions are relatively stable and commonly shared, which is due to the fact that we belong to the same biological species and to cultural and scientific communities with some shared knowledge. Nevertheless, intuitions are neither intrasubjectively nor intersubjectively absolutely stable, since they also depend on our changing experience and knowledge. This explains why the conclusion that all bodies fall at the same speed in Galileo's thought experiment seems to us (and to Brown) to be immediately clear, whereas Salviati in Galileo's *Discorsi* had a hard time getting Simplicio to reach this conclusion.

In the account I have sketched so far, we do not have to postulate a mysterious Platonic realm in which we "see" the right laws of nature and which is inaccessible to argumentation. On the other hand, we need not fly to the opposite extreme, according to which we actually execute argumentative reasoning processes while conducting thought experiments. Instead, there are some *a posteriori* acquired "truths" that function as implicit background knowledge, enabling us to come to a relatively quick decision in the evaluation of a thought experiment. But we can always make these premises explicit by reconstructing the thought experiment as an argument. This is useful, for example, when we are sceptical about the plausibility of a thought experiment. An argumentative reconstruction of a thought experiment can uncover implausible or false premises. It can show that important premises were overlooked and it can detect invalid inferences. By doing this we can critically reflect on those intuitions that emerge in the thought experiment and analyse whether they really represent justified beliefs or whether they lead us to wrong conclusions.

Like sense perception, introspection and memory, intuitions are an inevitable but fallible and sometimes unreliable source of evidence. Without trusting any of our intuitions, we would not be able to understand each other or the world. In a philosophical discussion about the legitimacy of thought experiments as a method of knowledge acquisition, we should not consequently reject all intuitions, but we should try to differentiate between classes of thought experiments which are legitimate and those in which misleading intuitions are involved.

4. The legitimacy of the method of thought experiments

Since, as we have seen, thought experiments can be construed as arguments, they can suffer from the same defects as unsound or unjustified arguments. Thought experiments can be dismissed because they are based on implausible, incoherent or inconsistent premises or because they involve inconclusive judgements, illogical inferences or other kinds of argumentative shortcomings like a *petitio principii*. However, thought experiments manifest those argumentative faults in characteristic ways. They often employ highly suggestive imaginary scenarios that appeal to intuitions and coerce them in certain directions. That is why some thought experiments give rise to particular intuitive conclusions where no rational and critical examination seems to be necessary. Dennett calls such thought experiments "intuition pumps" and describes them as follows:

> A popular strategy in philosophy is to construct a certain sort of thought experiment I call an *intuition pump* […]. Such thought experiments […] are *not* supposed to clothe strict arguments that prove conclusions from premises. Rather, their point is to entrain a family of imaginative reflections in the reader that ultimately yields not a formal conclusion but a dictate of "intuition". (Dennett 1984, 12)

In the following, I will try to single out some features that are typical of intuition pumps.

Even if thought experiments evoke imaginary scenarios, it is necessary that they still fulfil the theoretical requirements of an *experiment.* As mentioned above, these requirements involve the study and evaluation of a situation in which data is changed, modified or reorganized in a *planned* and *controlled* way with the help of specific *background hypotheses* or *theories.* What does this mean for the proper use of thought experiments? First of all, the imaginary scenario should not be under-determined in relevant aspects. In particular, this means that if we invent a scenario in which we manipulate or change data in an unfamiliar way, the effects of these manipulations or changes should always be under control, i.e., we should understand how they can affect other implicit assumptions of the thought experiment and whether these effects can still justify the intended conclusion of the thought experiment. Many thought experiments in philosophy suffer from such under-determination, which makes it hard to draw a conclusion since some relevant premises are unclear. Such thought experiments turn into illegitimate "intuition pumps" if the under-determination is cleverly disguised and the scenario is outlined in a way that leads intuitively to a conclusion which is not (or not completely) supported by the premises.

Putnam's famous *twin earth* thought experiment is a typical example of such an intuition pump. Putnam describes the twin earth as a planet which is exactly identical to the earth we live in except for the fact that the liquid in the rivers, lakes and seas of twin earth has the chemical structure XYZ which is different from H_2O. But nevertheless, in its surface structure this liquid cannot be distinguished from water on earth. It is further assumed that every person on earth has an exact "molecular copy" on twin earth. Our "twin earth Doppelgängers" also use the word "water" to refer to the liquid in their rivers, lakes etc. But they do not have any knowledge of the concept of H_2O. Putnam now argues that although there is no relevant difference between the mental states of our "twin earth Doppelgängers" and our own, the reference of the word "water" is different. Therefore, reference is not determined by psychological states. "Meanings are not in our heads".[12] I do not want to go into the details of the long discussion of this thought experiment and the question of whether semantic externalism is plausible or not. I just want to point out that by varying one particular factor of our world in his imagination (water is no longer H_2O but XYZ), Putnam fails to pay attention to the drastic effects this variation has for twin earth and its inhabitants. He merely, and illegitimately, stipulates that everything else remains the same. But of course, if the liquid on twin earth is not H_2O our "twin earth Doppelgängers" cannot be molecularly identical to us. About 70 % of a human being consists of H_2O molecules. If we exchange an important chemical substance with something else, the so-called twin earth will be completely different from the world we live in and – contrary to what Putnam will have us believe – we will have not the slightest idea of what this strange world and the psychological states of its inhabitants (if they have any) will be like. I know that similar objections to Putnam's thought experiment are frequently raised in the literature, but they are mostly considered as irrelevant or "beside the point". Maybe Putnam could make up

[12] See Putnam 1975.

another thought experiment with no such shortcomings. But nevertheless, his "twin earth" example is an illegitimate intuition pump. It does not fulfil the necessary requirement of a *controlled* change of data and the side-effects of this change remain completely under-determined. The implicit premise in this thought experiment, namely, that the mental states of the human beings and their "twin earth Doppelgängers" are identical even when the chemical or molecular composition of the two worlds are different, is not justified. It is merely stipulated without a supporting argument that such dramatic changes of the molecular structure will not cause any differences on the surface structure. That is why the conclusion drawn from this thought experiment rests on an illegitimate intuition pump.

David Ward coined the term "black box scenarios" to characterize thought experiments in which relevant background conditions are under-determined in such a way that we are unable to uncover a plausible explanation (for example by extrapolating principles we are already familiar with) of how the imaginary scenario could possibly be conceived.[13] So we can also analyse Putnam's twin earth example as a kind of a black box scenario since Putnam does not give us any explanation as to why the twin earth is not radically different from our world.

There is another reason why our intuitions can be misdirected. By contrast to the above mentioned under-determination, this kind of thought experiment involves an imaginary scenario that can be specifically described and embellished. Such thought experiments turn into intuition pumps, if we become distracted by embellishments in such a manner that we do not realize that the general conclusions drawn from the thought experiment are not justified by this *single* and *specific* example. Intuition pumps of this sort can be revealed by asking the question: "Does a structurally analogous example have the same intuitive plausibility and lead to the same consequences?"

A famous example of an intuition pump of this kind is Leibniz's thought experiment to refute a mechanistic approach of perception. In the *Monadology* he provides the following *reductio* argument:[14] Suppose perceptions (like thinking, feeling, perceiving) can be produced by a machine. Now imagine that this machine is enlarged in a way that one might enter it "as if it were a mill". For Leibniz it is quite obvious that if we were to enter this "mill" and look around, we would only observe mechanical processes ("parts which push and move each other"), but we would never observe anything that explains per-

[13] See Ward 1995.

[14] Leibniz 1714, Sec. 17.

ception. Consequently, perceptions cannot be explained in a mechanistic way, and the mind must be conceived as immaterial. This thought experiment (and similar thought experiments in the philosophy of mind in the 20th century aimed against a mere physicalist or functionalist explanation of mental states or qualia or against a mere formal, syntactic or computational explanation of meaning and understanding – like John Searle's *Chinese room*) faces the problem of its own strong persuasive power. We can easily imagine being in such a situation, and it seems intuitively clear that we cannot find anything that explains perceptions (or mental states or qualia) inside our imagined perception-producing mill (or inside a brain). In spite of its persuasiveness, this thought experiment is a question-begging intuition pump. It exploits the fact that when we are examining a complex phenomenon, we often find it intuitively implausible to account for properties observed at the subsystem level at the level of the whole system. This is particularly the case when the phenomenon in question is complex, abstract, and has not yet been completely explained by scientific inquiry. That is why we are inclined to reject the whole approach of explaining a system's phenomena by means of a subsystem's properties or at least to claim that there might be an explanatory gap. The concrete description of the imaginary scenario ("parts which push and move each other") helps to reinforce this inclination.

But we can easily find an example in which the explanation of a system's phenomenon by properties of the subsystem is much less problematic. Take, for example, the following argument given by David Cole:

> Imagine a drop of water expanded in size until each molecule is the size of a grindstone in a mill. If you walked through such a now mill-sized drop of water, you might see wondrous things but you would see nothing wet. But this hardly shows that water does not consist *solely* of H_2O. (Cole 1984, 432)

As long as we have no good reason – independent of the mere intuition-pumping scenario of Leibniz's thought experiment – for thinking that there is a fundamental structural difference between this argument and the original thought experiment, the conclusion Leibniz draws from his thought experiment is not legitimate and his thought experiment tends to be question-begging.

The counterfactual, fictional, or idealized assumptions made in many thought experiments can also be a source of fallacies. As Kathleen Wilkes has pointed out, it is necessary that the "impossibilities" involved in a thought experiment are irrelevant, i.e. that they are merely used in a heuristic way and do not nullify the intended goal of the thought experiment.[15] In order to recognize

[15] See Wilkes 1988, 9.

whether there are some relevant counterfactual assumptions involved, we have to specify the implicit background conditions of a thought experiment very carefully, and we have to check whether the set-up of the thought experiment is impossible according to those background conditions. For example, in one of Einstein's famous thought experiments, there is the counterfactual assumption that a person travels besides a beam of light. The aim of this thought experiment is to refute Maxwell's theory of electrodynamics. According to this theory, light is considered as an electromagnetic wave which spreads by the transformation from electric into magnetic field energy and from magnetic into electric field energy. If we suppose that someone travels besides a beam of light, the strength of the electric field for such a moving observer would be *constant.* The observer would be faced by the impossibility of a stationary oscillatory field. But this result contradicts Maxwell's theory according to which only a *changing* electric field could cause a magnetic field. So a necessary condition for the spreading of light would not be fulfilled.

Now compare this thought experiment with another that Einstein invented in order to refute Heisenberg's uncertainty principle in quantum mechanics. According to this principle, the energy of a photon and the exact time of the energy measurement of this photon cannot be measured simultaneously. Einstein imagined a box containing a clock. The box is fixed to a spring balance with a pointer that registers the movement of the box. The clock inside the box is connected to a shutter that opens or closes a small hole in the box. At an exact moment the shutter opens to release one single photon inside the box. We can therefore measure exactly the time when the photon is released. But, as Einstein argues, we can also measure exactly the energy of the photon. The difference between the weight of the box before and after the photon was released will provide us with the mass of the photon – and therefore by using the famous formula $E=mc^2$ we will get the energy of the photon. Contrary to the uncertainty principle, it seems that we can measure the exact time and energy of a photon simultaneously. Niels Bohr was not impressed by this thought experiment. Instead he defeated Einstein with his own weapons by pointing out that such a thought experiment contradicts principles of the general relativity theory. According to Einstein's theory of general relativity, the tempo of the clock is affected by its movements in the gravitational field. When the photon escapes, the spring balance will cause a minimal change of the clock in the gravitational field. Hence, the tempo of the clock involves a minimal uncertainty – which is predicted by Heisenberg's principle that Einstein actually wanted to refute.

Both of the above mentioned thought experiments involve counterfactual assumptions. In the first thought experiment it is assumed that travelling at the speed of light is possible. In the “clock in the box” thought experiment it is assumed that the measuring device for measuring simultaneously the time and energy of a photon is possible. This assumption is counterfactual when the theory of general relativity is taken into account. Now the following question arises: Why do we judge that Einstein’s first thought experiment was successful, but the second one failed? In the first thought experiment the counterfactual assumption is irrelevant since the possibility of a moving frame of reference at the speed of light is not excluded in Maxwell’s theory. Maxwell’s theory was intended as a *general* theory about the spreading of electromagnetic waves. Therefore, it is legitimate to ask what consequences Maxwell’s theory will have from the perspective of an observer moving at the speed of light. A “speed of light traveller” is neither explicitly excluded in Maxwell’s theory nor in any other background theory that is relevant for the execution of the thought experiment. In the second thought experiment, however, the counterfactual assumption is not irrelevant. The theory of general relativity functions as a background theory which renders the entire thought experiment inconceivable. The lesson to be learned from this is that we should try to make all the relevant background assumptions in a thought experiment explicit in order to be able to avoid the illegitimate employment of relevant counterfactual assumptions.

Finally, I would like to mention another important but problematic use of thought experiments where we also have to take care, in order not to be led astray by an illegitimate intuition pump. Sometimes philosophical thought experiments are used for conceptual analyses. These thought experiments confront us with situations in which we have to decide intuitively whether the given situation can be treated as a case of correct application of the concept in question. On my naturalistic account of intuitions outlined above, intuitions are to be viewed as mental propositional attitudes accompanied by a feeling of certainty. Although they seem to be spontaneous and non-inferential in character, they are not *a priori* in the sense of being irreducible or independent to empirically gained knowledge. Intuitions are shaped by our causal interactions with the environment. That is the reason why some intuitions concerning common concepts applied in familiar situations are relatively stable and why our intuitions diverge or fail when concepts are applied in a novel or unfamiliar setting. It also explains why intuitions can alter due to our changing experiences and knowledge. When we employ thought experiments in order to settle questions about the correct analyses of a concept, we have to take into consideration the fact that our intuitions may not help us in improbable and unfamiliar situations and that our judgements about the right application of the concept may well be distorted.

An example of a successful thought experiment that helps us to clear up the nature of a concept is the famous Gettier case.[16] Almost everybody agrees immediately that the situations Gettier presents are not cases of knowledge, although the person has a true and justified belief that *p*. For, in Gettier's examples, the justification of *p* rests on a false assumption, and *p* turns out to be true just by pure luck. Gettier's examples do not imagine fictional or unusual situations. We obviously have the very strong and stable intuition that knowledge is not mere justified belief that happens to be true by chance. So Gettier's thought experiments are not at all question-begging. They show how thought experiments can be legitimately employed in order to clarify an important philosophical concept. Through the help of some imaginary scenarios, Gettier showed successfully that the traditional definition of knowledge as true and justified belief is inadequate since truth and (a certain understanding of) justification are not sufficient for knowledge. His thought experiments also suggest how future analyses of knowledge could proceed: One could try to find a further condition for knowledge, one could try to specify the notion of justification in such a way that the Gettier-cases no longer count as justified beliefs, or one could give up the whole project of defining knowledge and instead restrict oneself to the analyses of important necessary conditions of knowledge.

In other thought experiments in philosophy where concepts are used in unfamiliar settings we are sometimes uncertain as to whether the application of a concept is legitimate or not. In particular, such thought experiments are widely used in the debate concerning the concept of personal identity when, for example, teletransportation, fission processes, brain transplantation or other mysterious methods from science fiction are involved. In such fictional situations our intuitions are of no great help, because it is very hard to decide in a non-question-begging way what the real criteria for personal identity are. That is why a lot of philosophers are very critical about the method of thought experiments in such applications. Wiggins, for example, points out that by "denaturing the human subject" in these thought experiments, we do not learn anything about the nature of personal identity because the decision whether the concept of identity is legitimately applied is more a question of stipulation and not a matter of discovery.[17] Another problematic case in such thought experiments is the frequently made implicit assumption that there exists *one* single, coherent concept of personal identity which adequately represents the real nature of identity. Different positions are therefore considered as competing

[16] See Gettier 1963.

[17] See Wiggins 1980, 178.

views about the unique correct understanding of identity. But if we disrupt a concept by applying it in unnatural situations, we may destroy the inner structure of the concept. So, if we take those unnatural situations seriously, it might be more plausible to conclude that there are two or more different concepts of identity.

For example, the long and sometimes fruitless debate in epistemology between internalist and externalist approaches to knowledge could indicate that there is not just one single concept of knowledge but at least two different concepts, each of which reflects different features of knowledge. One concept takes our externalist intuitions into account – for example, when we apply knowledge to small children or animals or when we use knowledge in situations where we gain a true belief by a spontaneous and reliable sense perception in normal circumstances. And the other concept of knowledge reflects our internalist intuitions – for example, when knowledge requires fulfilling our epistemic duty by intellectually reflecting on the justificational grounds of the belief. With the help of thought experiments these divergent, but legitimate concepts of knowledge can be clarified.[18] Furthermore, if an analysis or a definition of a concept is regarded as *universally valid,* then far-fetched imaginary scenarios that cast doubt on this general validity are legitimate. For example, this is very often the case in thought experiments in ethics where the universality of an ethical theory is rejected by counter-examples that normally do not occur in our everyday life (like Bernard Williams' "poor Jim" example in his *Critique of Utilitarianism*).[19]

Instead of condemning thought experiments as question-begging in situations where concepts are applied in unnatural novel settings (as, for example, Wiggins, Quine and Wittgenstein suggest), thought experiments may well be used in these situations in a theoretically useful way. They may help us to uncover borderline cases of the concept in question, or they may provide reasons for distinguishing between two different concepts.

Since intuitions are formed by interaction and adaptation to our environment and since they give the appearance of addressing abstract and general things or states of affairs, there always remains the risk of having fallible or oversimplified intuitions. They may be able to function as reasonably adequate "rules of thumb" for the application of concepts in normal cases but fail in specific unusual situations we have never thought about before. For example,

[18] Internalists and extermalists also seem to be employing different concepts of epistemic justification. See Engel 1992, for a discussion of several thought experiments that highlight these differences.

[19] See Williams 1973, ch. 3.

we normally have the strong intuition that the so-called "naïve" comprehension axiom or the T-schema ("p" is true iff p) holds in "naïve" semantics with a semantically closed language. But these intuitions are developed against the background of "normal" applications of set theory or semantics. A precise and deeper analysis of these principles has nevertheless shown that, when certain self-referential, i.e., "unusual" applications are taken into account, the principles lead to paradox. For mathematicians or philosophers who study the set-theoretic or semantic paradoxes intensively, these self-referential applications become quite normal, and therefore, their intuitions concerning these principles can change. Let us consider a philosopher working vigorously every day on the liar paradox, who as a result of her studies becomes convinced that Tarski's approach of dividing between an object- and a meta-language and rejecting the assumption of a semantically closed language is the correct solution to this paradox. When she is then confronted with the "naïve" T-schema, she will no longer have the intellectual disposition that this schema could be correct. Her intensive study of the liar paradox will indeed have an effect on her intuitions. So she will immediately intuit that the principle is wrong, since the self-referential applications leading to the paradox will automatically become apparent to her. Therefore, intuitions can be modified, changed or completely destroyed by our experiences.

Of course, there might also be other reactions to the liar paradox. For Graham Priest and other defenders of the so-called "dialethism", Tarski's approach is simply a technical and artificial solution to the liar paradox that is intuitively unconvincing. For Priest, the "naïve" T-schema is so deeply entrenched in our beliefs about truth that he will never be inclined to accept solutions in which the liar paradox might be explained away by technical dirty tricks. Instead, by insisting on his intuition that the "naïve" T-schema is correct, he is willing to believe that there are "true contradictions".[20] And the longer Priest works on the defence of his "dialethism", the more convinced he becomes about his position – and the more his intuitions concerning these facts become reinforced. So intuitions can vary from person to person as well as during the lifetime of a single person due to different experiences. Thought experiments by imagining situations in which our intuitions might go astray can reveal unstable, incoherent or even inconsistent intuitions and can therefore help to change or modify our intuitions about a concept. But they can also help to sharpen or strengthen our intuitions.

Let's take stock of what has been said about the legitimacy of thought experiments. Thought experiments are an indispensable method of argumenta-

[20] See, for example, Priest 1987 or Priest 1998.

tion in order to change, develop or strengthen theories or concepts. Therefore, we cannot manage without thought experiments. But they can be misleading in various ways. In particular, they can misuse intuitions and lead us to believe in an unjustified conclusion. Borrowing a term from Dennett, I called such cases "intuition pumps". I have tried to specify different ways in which a thought experiment can become an intuition pump.

In order to avoid intuition pumps, we must first of all pay attention to the fact that no relevant aspects of the thought experiment are under-determined, i.e., if we change data in an imaginary scenario, we have to know how they effect other things assumed in the thought experiment. and we have to be sure that these effects do not undermine the intended result of the thought experiment. In other words, we ought to avoid using "black box" explanations.

Second, the single case imagined in a thought experiment should justify the general conclusion that is the intended result of the thought experiment. The specific description of the scenario should not distract our intuitions in such a way that we fail to realize that a structurally analogous example will not have the same intuitive plausibility and will not lead to an analogous conclusion.

Third, counterfactual, fictional or idealized assumptions made in a thought experiment should always be irrelevant, i.e., the counterfactual, fictional or idealized assumptions should never be already excluded as impossible situations by the theory (or the background theories on which it relies on) that the thought experiment wants to investigate.

Fourth, thought experiments in which we are asked to apply a concept in an unfamiliar situation can also be problematic, since in far-fetched science-fiction cases, for example, we often have no stable intuitions that could guide us in finding a justified answer. So, as Wiggins has pointed out, it could become "a matter not of discovery but of interpretation (or even stipulation)". But the imagination of far-fetched situations in thought experiments should not be rejected wholesale, since they might cast light on borderline cases of a concept and unstable, incoherent or inconsistent intuitions and therefore help to clarify our intuitions and show different ways of finding an adequate analysis of a concept.

5. Conclusion

Although intuitions seem to be immediate and spontaneous propositional attitudes with a strong feeling of certainty, they are shaped by our experiences and by our adaptations to the environment. They are, like sense perceptions, fallible and unstable and run the risk of being systematically misleading (as some famous results of cognitive psychologists show). Nevertheless, they provide a definite source of evidence that we need in order to understand the world and to interact successfully with our environment. In particular, intuitions play an important role in the execution and interpretation of thought experiments which are, for their part, necessary tools for scientific inquiry. Contrary to Brown's opinion, there are no "Platonic" thought experiments in which intuitions provide us with *a priori* insights into a realm of abstract entities and universals. Instead, thought experiments can be entirely reconstructed as arguments, whereas intuitions are a kind of evidence that function as background knowledge which can be made explicit as premises in a thought experiment. Intuitions in thought experiments can be misdirected. The scenario imagined in a thought experiment can appeal to our intuitions in a way that illegitimately makes us believe in conclusions that are not really justified by the premises. So we have to employ thought experiments carefully. The criteria for a proper use of thought experiments explicated above will hopefully help us to be more aware of the pitfalls of intuition pumps and how to avoid them.

References

Brown, J. R. 1991a: *The Laboratory of the Mind,* London/New York: Routledge.

Brown, J. R. 1991b: "Thought Experiments: A Platonic Account", in Horowitz, T./Massey, G.J. (eds.): *Thought Experiments in Science and Philosophy,* Savage: Rowman & Littlefield.

Cole, D. 1984: "Thought and Thought Experiments", *Philosophical Studies* 45, 431-444.

Dennett, D. C. 1984: *Elbow Room,* Oxford: Oxford University Press.

Engel, Jr., M. 1992: "Personal and Doxastic Justification in Epistemology"*, Philosophical Studies* 67, 133-150.

Galilei, G. 1638: *Discorsi e dimostrazioni matematiche intorne a due nuove scienze attenti alla mecanica & i movimenti locali,* Leyden (English translation (1974): *The New Sciences,* Madison, WI: University of Wisconsin Press).

Gettier, E. L. 1963: "Is Justified True Belief Knowledge?", *Analysis* 23, 121-123.

Kuhn, T. 1977: "A Function for Thought Experiments", in Kuhn, T.: *The Essential Tension,* Chicago: University of Chicago Press.

Leibniz, G. W. 1714: *Monadologie* (English translation (1965): *Monadology,* Bobbs-Merrill).

Locke, J. 1690: *An Essay Concerning Human Understanding,* London.

Mach, E. 1897: "Über Gedankenexperimente", *Zeitschrift für physikalischen und chemischen Unterricht,* vol. 10, 1-5.

Mach, E. 1905: *Erkenntnis und Irrtum,* Leipzig, reprint ^{5}1926, 180-197.

Norton, J. D. 1996: "Are Thought Experiments Just What You Thought?", *Canadian Journal of Philosophy* 26, 333-366.

Ørsted, H. C. 1811: *Förste Indledning til den Almindelige Naturlaere, et indbydelsesskrivt til forelaesninger over denne vindenskab,* Copenhagen (German translation (1822): „Über Geist und Studium der allgemeinen Naturlehre", *Gehlens Journal für Chemie und Physik,* vol. 36, 458-488.)

Priest, G. 1987: *In Contradiction: A Study of the Transconsistent,* The Hague: Martinus Nijhoff.

Priest, G. 1998: *Stanford Internet Encyclopedia of Philosophy,* http://plato.stanford.edu/entries/dialethism/ .

PUTNAM, H. 1975: “The Meaning of Meaning”, in Putnam, H.: *Philosophical Papers,* Vol. 2, Cambridge: Cambridge University Press.

SORENSEN, R. A. 1992: *Thought Experiments,* Oxford: Oxford University Press.

WARD, D. 1995: “Imaginary Scenarios, Black Boxes and Philosophical Method”, *Erkenntnis 43,* 181-198.

WIGGINS, D. 1980: *Sameness and Substance,* Oxford: Basil Blackwell.

WILKES, K. 1988: *Real People: Personal Identity Without Thought Experiments,* Oxford: Clarendon Press.

WILLIAMS, B. 1973: “A Critique of Utilitarianism”, in Smart, J.J.C./Williams, B.: *Utilitarianism. For and Against,* Cambridge: Cambridge University Press.