# Thought Experiments

## Rachel Cooper

**Abstract**

This paper seeks to explain how thought experiments work, and also the reasons why they can fail. The paper is split into four sections. The first argues that thought experiments in philosophy and science should be treated together. The second examines existing accounts of thought experiments, and shows why they are inadequate. The third proposes a better account of thought experiments. According to this account, a thought experimenter manipulates her world view in accord with the "what if" questions posed by a thought experiment. When all necessary manipulations are carried through the result is either a consistent model, or contradiction. If a consistent model is achieved the thought experimenter can conclude that the scenario is possible, if a consistent model

cannot be constructed then the scenario is not possible. The fourth section of the paper uses this account to shed light on the circumstances in which thought experiments fail.

**Keywords:** Thought experiment, Thomas Kuhn, John Norton, James Brown.

This paper seeks to provide an account of how thought experiments work, and of how they can go wrong. Philosophers should be interested in this project for two reasons. First, philosophers often use thought experiments, especially in ethics and the philosophy of mind, and an understanding of how thought experiments work might enable philosophers to use them more successfully. Second, thought experiments are epistemically interesting in their own right. In a thought experiment it seems we can start from a position of ignorance, sit and think, and gain new knowledge, despite the input of no new empirical data. One aim of this paper is to explain the origin of this new knowledge.

The paper is split into four sections. The first argues that thought experiments in philosophy and science can be treated together. The second examines existing accounts of thought experiments and shows why they are inadequate. The third proposes a better account of thought experiments. The fourth uses this account to shed light on the circumstances in which thought experiments fail.

Before the philosophical work, it will be useful to clarify what I mean by "thought experiment". For the purposes of this paper I shall adapt a definition offered by Tamar Szabó Gendler and take it that to conduct a thought experiment is to make a judgement about what would be the case if the particular state of affairs described in some imaginary scenario were actual (Gendler, 1998, 398).

## 1.Thought experiments in science and philosophy

Philosophers writing on thought experiments divide between those who restrict their attention to thought experiments in science, and those who consider thought experiments in both philosophy and in science together.[1] Those who only consider thought experiments in science have not given arguments for thinking that thought experiments in science are necessarily different. Their restriction seems to result from a strategy of caution – these authors are not sure whether thought experiments are similar in philosophy and science, and so just talk about areas where they are convinced their account works. In this paper I throw caution to the wind, and concern myself with thought experiments in all areas. There are two reasons why I think this is the best way to proceed. First, on grounds of simplicity, if it is possible to produce a unified account of thought experimentation this should be preferred. And, the only way to find out whether there is an acceptable unified account is to try and construct one. This is what I attempt in this paper.

Second, there are reasons to be sceptical of the idea that science and philosophy are radically distinct enterprises. The work of empirically-inclined philosophers of mind and language is often indistinguishable from work in theoretical psychology or linguistics. The same holds for philosophers of physics and theoretical physicists, and for game theorists, economists, and theoretical evolutionary biologists. In many cases, philosophical and scientific pieces of work can only be distinguished on the basis of the journals in which they are published. Moreover, because of its necessarily non-

empirical nature, work involving thought experiments is particularly likely to fall on the border between philosophy and science. Articles on the E. P.R experiment, Schrödinger's cat, or bilking experiments (thought experiments showing that causal paradoxes would emerge if one could go back in time and kill one's father), are as likely to be found in the Physical Review as the Philosophical Review. Newcombe's paradox is discussed equally by economists and philosophers. Psychologists and philosophers alike worry about the Turing Test and Searle's Chinese Room. It is hard to distinguish science from philosophy, and even harder to distinguish philosophical from scientific thought experiments. For this reason an account of thought experimentation that can encompass all thought experiments, whether "philosophical" or "scientific", is to be preferred.

Occasionally, it has been suggested that while we may not be able to divide thought experiments into philosophical and scientific, they can be divided into distinct classes on the basis of the type of question that they ask. Some thought experiments, it is said, ask what would happen in a hypothetical state of affairs, others ask how we would describe situations, and yet others, how we would evaluate them. While I accept that thought experiments can be employed to answer different types of question, I suggest that it is a mistake to think of there being a corresponding variety of different types of thought experiment. This is because it is implausible to think that there are distinct mental processes at work in considering how things are, in describing them, and in evaluating them. Our imaginings are shaped by how we describe situations, and many descriptions are already value-laden. Thus, if I imagine a small boy setting fire to a cat, I do not first form the image, and then label it a case of torture, and then decide that it wouldn't be very nice of the boy. Rather I imagine a cruel boy torturing a cat – the

description, and evaluation, are already built into the hypothetical scene. As we describe, and evaluate, alongside imagining, there will not be different types of thought experiment, some of which just involve imagining, and others of which involve additional activities.

## 2. Existing accounts of thought experiments

In this section I examine accounts of thought experiments have been proposed by other authors, and show why they are inadequate.

### a. Kuhn's remembering account

Thomas Kuhn puts forward an account of scientific thought experiments in his 1964 paper, "A function for thought experiments". According to Kuhn, during periods of normal science, scientists see anomalies but typically turn a blind eye to them. Usually a scientist's experiences of anomalies quickly fade from memory. Knowledge of anomalies is not necessarily altogether lost, however, as certain techniques can be employed to bring this semi-forgotten knowledge back into consciousness.

In Kuhn's account, thought experiments work by providing scientists with a means of retrieving memories of anomalies that they have previously seen, but so far ignored. The narrative structure of a thought experiment acts to trigger the memory of the scientist. As a scientist visualises the scenario sketched by the thought experiment, he experiences a feeling of déjà vu. This is because he has seen the scenario before, and so when prompted by the structure of the thought experiment he can work out what

would happen if the imagined scenario were actual. Kuhn argues that thought experiments have an important role to play in the history of science. They enable repressed knowledge of anomalies to come to the attention of scientists. This enables them to appreciate when their paradigm is inadequate, and can thus help normal science to enter a revolutionary phase. Kuhn's account manages to explain how knowledge can be gained via thought experimentation. The "new" knowledge gained in a thought experiment is remembered knowledge. As such it is not really new, and the epistemic puzzle of how armchair experiments can yield knowledge is solved.

Kuhn himself accepts that his account may not apply to all thought experiments, and indeed there is a large class of thought experiments for which it cannot account. According to Kuhn, a scientist can work out what would happen in a hypothetical situation because he has seen situations of the type being described in the real world. The scientist just has to remember what he has seen previously. However, some thought experiments concern situations that cannot have been seen before. Consider thought experiments that involve physically impossible scenarios. Einstein running along a light beam, for example, or Poincaré's Flat Land, which involves 2D people exploring a 2D environment (Poincaré, 1952, 37-8). As physically impossible situations cannot have been perceived previously, Kuhn's account is incapable of coping with such thought experiments.

Simplicity dictates that a common account of all thought experiments should be sought if at all possible. For this reason, if an account can be given that encompasses physically impossible thought experiments along with others, this should be preferred to Kuhn's. Later in this paper I propose such an account.

**b. Norton's argument account**

John Norton has proposed an account according to which thought experiments are really just dressed-up arguments (Norton 1991, 1996).[2] He aims principally to give an account of thought experiments in physics, and supports his position by offering reconstructions of the formal arguments that he thinks underlie some of Einstein's thought experiments. Norton fails to specify precisely what he means by "argument". However, he cannot simply mean "deductive argument", as he explicitly accepts that thought experiments can use inductive as well as deductive inferences. Norton's general approach in showing that thought experiments are arguments is to reduce them to a series of propositions. He reduces them to lists of premises and assumptions, leading to a conclusion via inferences of a recognised sort. This, I shall take it, is what at minimum he means by "argument". Norton claims that all thought experiments can be reduced to such arguments without epistemic loss.

Norton's account should not be accepted.[3] His primary reason for thinking that thought experiments are arguments is that he has shown that some of Einstein's thought experiments can be replaced by arguments, but this demonstration is not sufficient to prove his claim. All Norton has shown is that Einstein's thought experiments lead to a conclusion that can also be reached via a logical argument. This is insufficient to demonstrate that the argument and the thought experiment are actually identical, as the processes via which the conclusion is reached may be quite different in the two cases. Indeed the phenomenology of thought experimentation suggests that this is the case. Simply put, constructing a thought experiment feels quite different from producing a

logical argument. Thought experiments are often fun and easy, arguments are usually not. When we perform a thought experiment we imagine the situation unfolding in our mind's eye. We don't consider premises, modes of inference, and conclusions.

Furthermore, in some cases, thought experiments require types of reasoning that cannot be considered argumentative in any sense, that is that cannot be reduced to anything like a premises-conclusion form. Take Hume's missing shade of blue (Hume, 1978, 6). Hume asks us to consider whether someone could imagine what the missing shade of blue looked like without ever having seen it. How do we perform such a thought experiment? I suggest that we do something like the following: we consider something like the colour charts for shades of paint and imagine a gap, and then we try and imagine the missing shade. This thought experiment requires us to imagine what it is like to see blue, something that cannot be reduced to propositional form. Other thought experiments that involve imagining qualia will similarly not be reducible to arguments, nor will thought experiments that require spatial reasoning, for example, one in which we see that a square peg cannot go through a round hole of the same diameter. Whatever thought experiments are, they're not simply arguments. Thus Norton's account must be rejected.

**c. Brown's Platonic account.**

James Brown agrees with Norton that some thought experiments are merely dressed up reductio arguments (Brown, 1991a., 76). However, he agrees with me that this cannot be the full story, and that some thought experiments are not arguments. To account for

these, in his 1991 book, The Laboratory of the Mind, Brown proposes a Platonic account of thought experimentation, which he models on Platonic accounts of mathematics.[4] According to Platonic accounts of mathematics, mathematical knowledge can be gained via perceiving, or intuiting, a Platonic realm of numbers. Brown claims that the laws of nature are relations between universals, and that thought experiments enable us to gain new knowledge of the laws of nature by providing us with access to a Platonic realm. When, for example, a physicist constructs a thought experiment concerning the behaviour of masses, the physicist gains knowledge via directly perceiving the relations between Platonic universals of masses. Brown is concerned primarily with thought experiments in science, but he suspects his account will work for philosophical thought experiments too.

There are several problems with Brown's account. First, there is no account of how the Platonic universals are "perceived". Brown attempts to block this objection by claiming that the mechanisms whereby physical objects are perceived are also poorly understood. Here he misses the force of the objection, which is derived from a causal theory of knowledge. A causal theory of knowledge holds that a necessary condition for knowledge is that a causal chain links us to the situation we claim to know about. Once we combine a causal theory of knowledge with a claim that causes must be physical, or at least spatio-temporal, we rule out the possibility of gaining knowledge of Platonic universals. [5] In addition, if a causal theory of reference is adopted, then a parallel argument shows that reference could not be made to any Platonic universals that might exist.

These arguments against Brown are not fully persuasive, however. Brown will almost certainly argue that these causal accounts run into problems in the case of

mathematics, and must thus be rejected. Here I think we reach an impasse. Debates over the nature of mathematical knowledge are too long running for there to be much hope that they will be resolved any time soon. In a well-known article Paul Benacerraf (1973) has argued that all current accounts of mathematics are unsatisfactory. Some accounts have concentrated on providing a satisfactory account of mathematical truth, but run into problems when explaining how we can come to know about these truths. Here Platonic accounts are the primary examples. Other accounts, such as the various forms of formalism, can deal with the epistemology of mathematics, but fail to provide a satisfying account of mathematical truth. As such, in the current state of play, Platonic accounts of mathematics appear unsatisfactory, but their defenders can rightly point out that they can deal with problems that other accounts of mathematics currently can not. In the absence of an account that is clearly better, Platonic accounts of mathematics cannot be entirely ruled out.

This being said, it is worth noting that a Platonic account of thought experiments requires a metaphysics even more bountiful than that required by a Platonic account of mathematics. A Platonic account of mathematics just requires there to be Platonic mathematical objects. Brown needs a far richer Platonic realm. He needs Platonic universals corresponding to Newton's rotating bucket, and to the string that ties Galileo's masses together, for example. Brown might protest at this and claim that he needs only universals that correspond to the basic physical laws – thus there will be universals of Mass and Force and F=ma, but no Bucket or String. Granted, the thought experimenter accesses the realm of the Fundamental Laws of Nature through telling a story about tied masses, Brown might say, but once they have achieved access, they perceive Mass and Gravity, rather than String. This will not do, however, as it does not

tie in with the phenomenology of thought experimentation. Brown might claim that the phenomenology is misleading. But, if he makes this move, his account is considerably weakened. One of the advantages claimed for his account was that it would explain the pseudo-visual nature of thought experimentation, and the ease and assurance with which conclusions can sometimes be drawn from thought experiments. If Brown claims that the phenomenology misleads, then he can no longer claim these advantages for his account.

To sum up: While there may not be anything better currently available, Platonic accounts of mathematics appear unpromising. Brown's account of thought experiments shares the problems of such accounts. In addition, even if it were possible to perceive universals, Brown's Platonic heaven would need to be repulsively over-populated. For these reasons his account should be countenanced only as a very last resort.

**d. Experimentalist accounts**

Some authors claim that thought experiments are literally experiments (Sorensen 1992a., Gooding 1990, McAllister 1996). They accept that regular experimenters manipulate the world, while thought experimenters manipulate thoughts, but think that this difference is insignificant compared to the features thought experiments and real experiments have in common. For example, both real and thought experiments can be used to demonstrate the inadequacies of theories, both involve isolating features of phenomenon that are of interest, and so on.

I'm not sure what to make of claims that thought experiments are literally

experiments. It's not as if experiments form a natural kind, such that it might be discovered that thought experiments are a species of the genus. Rather than being a claim like "Whales are mammals", the claim that thought experiments are experiments seems more like "Beanbags are chairs". Beanbags are like chairs in some respects, and someone who claims that beanbags are chairs seeks to direct our attention to these common features. Still, it remains the case that beanbags and chairs have important differences. Similarly, thought experiments are similar to real experiments in some ways, and not in others. When shorn of its rhetorical effect, the claim that "Thought experiments are experiments" comes down to no more than the claim that studying the similarities between thought experiments and real experiments is enlightening. This may well be the case. However, the fact that real experiments involve manipulations on material objects, while thought experiments do not, is a difference between thought experiments and real experiments that cannot be ignored. Real experiments can teach us about the world because they involve interacting with the world. In contrast, thought experiments are problematic because the source of the knowledge gained via thought experiment is unclear. Crucially, claiming that thought experiments are real experiments does not help explain the source of the knowledge gained via thought experiments.

When it comes to explaining how we can learn from thought experiments, those authors who claim that thought experiments are literally experiments supplement their account in various ways. In addition to claiming that thought experiments are experiments, Sorensen holds that thought experiments are paradoxes. They correspond to "a set of individually plausible yet inconsistent propositions" (Sorensen, 1992a., 6). In so far as a thought experiment is identified with a set of propositions, however,

Sorensen's account will run into the same kinds of problems as Norton's argument-based account. There are some thought experiments that simply do not have a propositional form. David Gooding says that thought experiments involve the "construction of experimental narratives that enable virtual or vicarious witnessing" and that "thought experiments work because they are distillations of practice" (Gooding, 1990, 204-205). Unfortunately, Gooding doesn't elaborate further, but in these comments he may be edging towards a model-based account of thought experimentation similar to that outlined in the next section.

## 2. A better account of thought experimentation

In this section I propose an account that explains thought experiments as attempts to construct models of possible worlds. Nancy Nersessian (1992) and Nenad Miščević (1992) have also proposed model-based accounts of thought experimentation. Their accounts differ substantially from my own in ways that will be spelt out later. In addition, and as mentioned previously, David Gooding (1990) makes some comments that suggest he holds some kind of model-based account, and Kathleen Wilkes (1988) talks of thought experimenters imagining possible worlds.

Characteristically, thought experiments present us with a series of "What if" questions. For example, we may seek to discover what would happen if there were no friction, or what would happen if people split like amoeba. In performing a thought experiment we temporarily adjust our world view in order to construct a model in accord with the answers to these "what if" questions.

When answering the "what if" questions we predict how imaginary entities would behave in the same way that we predict how real entities will behave. Sometimes we will have explicit laws governing how entities of the type we are imagining act in the types of situation we are imagining. Thus, we can predict how fast imaginary masses would fall under gravity in the same way that we can predict how fast real masses fall. We plug the relevant values into equations and calculate the prediction. Whether the masses are real or imaginary makes no difference. We can also employ tacit understanding of laws that we could not formally state. Sometimes the answers to the "what if" questions are provided by implicit laws that are contained in the implications of the concepts we are employing. For example it is part of the meaning of "light" that it travels at the speed of light, and part of the meaning of "pencil" that it is a writing implement. We can also employ simulation type reasoning. If the simulation account of out folk psychological practices is correct (as proposed by Gordan, 1986), then this type of reasoning would be employed to predict the behaviour of imagined people.

Roy Sorensen, and before him Ernst Mach, suggest that evolution has fitted us with modal intuitions that can be expected to be broadly accurate, at least within commonplace domains (Sorensen 1992a., 1992b.; Mach 1960). The idea, loosely, is that those of our ancestors who correctly intuited how lions behave in nearby possible worlds were better able to outwit the lions and stay alive. Believing that lions can jump 10 ft but not 100ft, that lions are killed when large rocks fall on them, and that if a lion catches you she'll eat you, had survival value. Against Sorensen, James Maffie (1997) convincingly argues that we should expect any evolved ability to intuit modal properties to be limited. "For what difference does believing '2+2=4' is necessarily vs.

nomologically vs. universally yet accidentally true make in terms of an organism's evolutionary fitness?" (Maffie, 1997, 213). Still, this is not sufficient to rule out the possibility that we might have evolved reliable intuitions regarding practical possibility, and if this were so, such instincts could also be used working out what would happen in hypothetical situations.

A point of key importance to my account is that the reasoning employed in constructing thought experiments is of a perfectly commonplace kind. Answering the "what if" questions of a thought experiment uses the same kind of processes as answering "what if" questions in all other contexts. As human beings, planning, plotting, and imagining are of great importance to us. If, for example, we are going to decorate a room or book a holiday we don't just go ahead and do these things, but instead spend some time considering the different courses of action available to us. We consider what would happen if we went on holiday to Bournemouth as compared to if we went to Turkey, and by comparing the anticipated consequences come to a conclusion as to which type of a holiday we would prefer. The forms of reasoning involved in such planning are identical to those involved in thought experimentation.

When a thought experimenter is faced with a "what if" question, she attempts to answer it in a rigorous fashion. She follows though all the relevant implications of altering one part of her world view and attempts to construct a coherent model of the situation she is imagining. The rigour with which thought experimenters attempt to answer "what if" questions is what differentiates thought experiments from daydreams and much fiction.[6] In a day dream I might lazily imagine being Prime Minister – there I am bossing everyone about, issuing edicts that extend university vacations, and so on. In a thought experiment such slap-dash imaginings are not permitted. If I conduct a

thought experiment in which I dictate that university vacations should be extended, then I am obligated to at least sketch a coherent model of the situation – the courses must be correspondingly shorter, degrees must be longer, funding per a student greater, and so on.

The thought experimenter is committed to rigorously considering all relevant consequences in answering the "what if" questions. Some consequences, however, will not be relevant to the purpose of the thought experiment and can safely be ignored. Consider, for example, the thought experiment in which Einstein considered what he would see if he ran along a light beam at the speed of light. Now, of course, anyone running at such speeds would be in no position to make observations: long before reaching light speeds they would be too tired to notice anything, and their running shoes would burn up. Such points, however, are irrelevant to the issues at hand and so can be ignored.

When the thought experimenter has followed through all relevant consequences of the "what if" questions, several outcomes are possible. Sometimes when all the "what if" questions are answered, the result is an internally consistent model. What do I mean by model? A dynamic representation of a situation. The model might consist of a set of propositions describing a situation, or it might be pictorial. In my view the form of the model may well differ in different cases, and doesn't much matter. Indeed I would go so far as to claim that whether a situation is modelled in thought alone, or in some more concrete medium such as plasticine, isn't all that important. Human beings have developed their capacity to think via utilising various aids – pen and paper, diagrams, and so on. In a sense such tools enable us to externalise thinking. In many cases the same mental operations can be performed in different ways. Consider, for example,

doing sums in mathematics. Some people can only do sums on paper. Some need to use their fingers and toes, and count them all up to find the result. Other people can do maths in their heads – of these some will imagine what the sum would look like if written down, while others use different methods. All these individuals are adding up, and it doesn't much matter what method they use. I suggest that the differences between mental models and concrete models can be similarly insignificant. One thought experimenter will be able to visualise a situation, another will use a scrawled diagram, and a third will need to use concrete objects to represent the actors. All three model the situation, and the differences between them are unimportant.

Depending on the account of possible worlds adopted, when the thought experimenter produces an internally consistent model she either constructs or represents a possible world. Adopting a realist stance towards possible worlds commits one to providing some explanation of how we come to know about these "other worlds", leading to difficulties similar to those that I considered problematic for Brown's Platonist account of thought experiments. Thus, here I will adopt an anti-realist account of possible worlds. This is not an essential element of my account, however, and those who are willing to countenance realist accounts of possible worlds can consistently also accept my account of thought experiments.

Strictly speaking, as the thought experimenter will not specify irrelevant details in her model, she will not produce a single possible world, but rather a template for an infinite number of possible worlds. "Possible world", in the singular, can be taken throughout as shorthand for this infinite set. If the thought experimenter manages to construct an internally consistent model, and thus construct a possible world, then she can conclude that the situation she has imagined is possible. The strength of the

possibility, physical or logical, depends on whether the thought experimenter has constrained herself to constructing only models where the actual physical laws obtain.

In some cases, the thought experimenter will be forced to conclude that an internally consistent model cannot be produced. Often this will be because following through the "what if" questions would result in a contradiction. In other cases, although there is no overt contradiction, the thought experimenter will conclude that an internally consistent model cannot be produced after numerous attempts to construct such a model have failed. In these cases different parts of the model simply will not go together, in a sense analogous to the sense in which the pieces of jumbled jigsaw puzzles cannot be made to fit together. If the thought experimenter decides that no internally consistent model can be produced she will conclude that the hypothesised situation is impossible. Again the strength of the impossibility depends on whether the thought experimenter has restricted herself to attempting to construct models in which the actual physical laws obtain.

Many regard inferences from "It is conceivable that X" to "It is possible that X" with suspicion (See, for example, Wilkes, 1988, 17). The claim that because I can form a picture of a fire-breathing dragon in my mind, fire-breathing dragons are possible is indeed dubious. However, my model-based account of thought experiments avoids these problems. The thought experimenter does not simply visualise herself dropping linked masses, for example, rather she constructs a model in which she drops linked masses using what she knows about physical laws and the implications of her concepts. Physical laws, and our concepts, have modal implications built into them already. The law that masses attract each other implies that masses in all physically possible worlds attract. Similarly, our concept of number implies that whatever other scandals may one

day come to light the number five cannot be the illegitimate offspring of Tony Blair (example adapted from Nagel 1998). In so far as we believe our scientific theories to be correct, and have a good grasp of our concepts, we can use them to support modal claims. Thought experiments merely make use of modal implications to which we are already implicitly committed.

On my account, thought experiments can show us whether or not a situation is possible. In doing this they can indirectly teach us about the actual world. Discovering that a situation is impossible shows us how the world cannot be. Similarly, discovering that a situation is necessary shows us how the world must be.

Thought experiments can also be used to explore our model of the actual world, that is they can be used to reveal the implicit consequences of our theories about the world. Thought experiments that seek to discover what our intuitions would be in hypothetical circumstances are of this type. In such thought experiments we construct, and in the process, describe, a possible world in which there are apparently intelligent Martians, or in which someone is presented with the option of killing one person to save many. Such thought experiments teach us nothing about the world, but rather allow us to explore the implicit consequences of our pre-existing beliefs.

Brown has claimed that thought experiments can also provide us with new knowledge about what contingently happens to be the case in the actual world. If Brown's claim is correct such thought experiments pose a serious threat to my account, as it is difficult to see how knowledge of contingent states of affairs can be derived from the construction or representation of possible worlds.

Brown's putative example of a thought experiment that provides us with knowledge of contingent matters of fact is one that Galileo used to both demonstrate the

falsity of Aristotelian physics, that held that heavy bodies fall faster than light bodies, and also to suggest the correct Galilean result, that all bodies fall at the same rate (Galileo, 1974, 66-67). Galileo asks us to imagine two falling bodies, one heavy and one light, that are tied together. The Aristotelian principle leads us to conflicting conclusions. First, we can conclude that, since the light body falls more slowly, by tying the two together the heavy body will be slowed down. Second, we can conclude that since the mass of the compound body is greater than that of the heavy body alone, the heavy body will now fall more quickly. This contradiction shows that Aristotelian physics is wrong and that heavier bodies cannot fall more quickly that light bodies. This is the reductio stage of the thought experiment. So far my account has no problems, I can say that the thought experiment showed that no consistent model could be produced in which heavy bodies fall faster than light bodies, and that thus it could be concluded that the Aristotelian scenario is impossible and so not true of the actual world.

The second stage of the thought experiment is more problematic. Galileo now goes on to draw the conclusion that all masses fall at the same rate. An important point is that Galileo need not have reached this conclusion. Showing that heavy bodies do not fall faster than light bodies is consistent with a multitude of alternative theories, such that red balls fall faster than balls of other colours, that square objects fall faster and so on. Brown thinks that Galileo's success in picking the right theory can only be explained by his Platonic account. For Brown the thought experiment enables Galileo to perceive the Platonic laws that govern the movement of masses and so see that all masses fall at the same rate.

The challenge for my account is to explain how Galileo could have gained knowledge of contingent states of affairs though constructing or representing possible

worlds. I think that Galileo's success can be accounted for by thinking of the thought experiment as one that shows that a situation is impossible, working in tandem with various background assumptions. The background assumptions are that colour, shape, chemical composition and so on have no effect on the rate at which a mass falls. These background assumptions serve to limit the options available to Galileo as he attempts to discover the laws governing the behaviour of falling masses. The only options consistent with the background assumptions are that heavy masses fall more quickly than light masses, that light bodies fall more quickly than heavy bodies, or that all masses fall at the same rate. The reductio stage of the thought experiment shows that heavier masses cannot fall more rapidly than light masses, and a parallel thought experiment would show that light bodies cannot fall more quickly than heavy bodies. Thus the thought experiment can reveal that the remaining option, that all masses fall at the same rate, is correct. However, this option is not generated by the thought experiment as Brown mistakenly believes, but was put into the thought experiment at the beginning as a background assumption. My modelling account can allow for knowledge of contingent states of affairs that is generated in such a way, and so escapes the threat posed by Brown's claim that thought experiments can teach us about contingent matter of fact.

The account of thought experimentation I have put forward in this section is similar in some respects to those proposed by Nancy Nersessian (1992) and by Nenad Miščević (1992). The main claim of all three accounts is that a thought experimenter gains knowledge through manipulating a model. There are however, important differences between the other accounts and my own.

First, and most importantly, Nersessian's and Miščević's models are specifically mental models of the type thought by some cognitive psychologists to be involved in the comprehension of narratives. Specifically, both philosophers claim to have based their accounts on the work of the psychologist P.N.Johnson-Laird. Nersessian tells us, rather mysteriously, that a mental model is not a linguistic representation, nor a picture in the mind, but a "structural analog of the situation described" (Nersessian, 1992, 297). Miščević seems to have a more pictorial view of mental models and claims that mental models have a "concrete and quasi-spatial character" (Miščević, 1992, 220). Both accounts are based on contestable empirical data. If it turns out that mental models of the type posited do not exist, then these accounts must be rejected. In contrast, my account uses a much looser notion of "model". Whether the thought experimenter reasons through the situation via manipulating a set of propositions, or a mental picture, or even plasticine characters, makes no difference to my account. In my account the form of the model is unconstrained. This means that my account can cope with possible changes in the details of psychological theory in a way that Nersessian's and Miščević's cannot.

Second, Nersessian's models are restricted to simulating the way in which phenomena would unfold in the real world (Nersessian, 1992, 295). Miščević's examples suggest that he holds a similar view. In contrast, in my account, modelled phenomena do not necessarily unfold as they would in the real world as the thought experimenter may model a world in which some laws of nature are suspended or altered. This difference is important. It means that my account can cope with thought experiments that hypothesise physically impossible situations.

Third, Nersessian's models are manipulated in accord with a special "simulative model-based reasoning" (Nersessian, 1992, 296). This reasoning specifically excludes the use of deductive and inductive inferences, as it is not performed on propositions (Nersessian, 1992, 297). In my account the basic forms of reasoning used to manipulate the model will be the same as those we use to predict occurrences in the real world: although such reasoning is not limited to induction and deduction, such inferences are definitely permitted. On this point Miščević agrees with me. He also allows that deductive and inductive reasoning can be employed (Miščević, 1992, 215).

## 4. When thought experiments fail.

My account predicts that thought experiments may fail in two ways. The first reason thought experiments can fail is if the thought experimenter is unable to answer the "What if?" questions correctly. Maybe she has no knowledge, either explicit or implicit, of the laws that govern the behaviour of the type of entities she is imagining. Maybe she has knowledge of the laws relevant for predicting the behaviour of entities of the imagined type in the actual world, but the laws do not apply in the hypothesised situation.

I suggest that Bernard Williams' thought experiment concerning people that split like amoeba is an example of a thought experiment that fails because we are unable to answer the necessary "what if" questions (Williams, 1973, 23). Williams asks, "What if people split like amoeba?", but we are unable to answer. How exactly could people split like amoeba? Would they split down the middle, and have one leg and one hand

each? In that case they would fall over, and unless skin suddenly grew to cover their wounds their organs would fall out. Or are they supposed to split into two mini but complete people? Then presumably, prior to splitting, a person would have to sprout an extra head, legs and arms. Either way, the biological logistics required to get the scenario off the ground are too complex and gruesome to work out.

The thought experimenter is more likely to make a mistake in answering the "what if?" questions if the laws she is using to provide the answers are implicit rather than explicit. When a law is explicit, as they typically are in field such as physics, the thought experimenter can clearly see whether the law applies to the situation she is imagining and knows how to apply it. In areas where the concepts we use are less well defined the thought experimenter has to sharpen her concepts as she goes along. As she searches for ways of extending her concepts to deal with previously unencountered circumstances she will often rely on analogy. However, reasoning by analogy depends on perceived similarity and what similarities are perceived is critically influenced by the way in which two situations are presented. This is why thought experiments using vaguely defined concepts, for example "person" are so open to criticism (see Wilkes, 1988, for more on this point).

The second reason why thought experiments can fail is because the thought experimenter can make a mistake as to whether she has constructed an internally consistent model. Inconsistency may be difficult to spot. Mathematicians can construct superficially convincing but false proofs, and Escher's pictures appear to represent actually impossible situations.

As an example of a thought experiment that fails because the thought experiment thought there was inconsistency where there is not, consider this ancient

thought experiment that purports to show that the universe is infinite by a reductio argument (taken from Sorensen, 1992a, 115). We are asked to imagine a man at the end of the universe who throws a spear. The spear cannot go forwards because there is literally nowhere for it to go. Thus it must rebound, which is absurd. We are left to conclude that the universe has no edge and so is infinite. This thought experiment fails because the thought experimenter has overlooked the fact that it is actually possible for a surface to both be finite and have no edge, the surface of a sphere is an example. The thought experimenter mistakenly saw a contradiction when there is none.

In general, we can say that a thought experiment is more likely to succeed if the thought experimenter is knowledgeable about the relevant aspects of the actual world. Only if she possesses either explicit or implicit knowledge of the behaviour of real phenomena can the thought experimenter predict how hypothetical events would unfold. It also helps if the knowledge being used to answer the "what if" questions is explicit rather than tacit. When our knowledge of a law is explicit, as it typically will be in fields such as physics, we can see clearly whether the law applies to the situation being imagined. A final, and rather mundane, thought is that it will best to keep the imagined situation as simple as possible if we are to avoid getting confused as to whether or not an imagined scenario is consistent.

Thought experiments can lead us astray. This has led some to suggest that they should be abandoned, and that thought experiments can and should be replaced by real experiments.[7] These writers are wrong. Although fallible, thought experiments are required for several reasons. Some thought experiments are practically possible, but there may be sound reasons for performing them in thought only. They may be unethical, or far too expensive to perform in practice. Other thought experiments cannot

be replaced by real experiments because they are physically impossible. These thought experiments may either involve idealisation or the direct violation of physical laws. I shall argue that it is untenable to reject the use of thought experiments of either type.

First, for thought experiments involving idealisation. An example is Galileo's thought experiment demonstrating that bodies continue moving with constant velocity in the absence of a force (as described in Sorensen, 1992a, 8-9). Galileo asks us to consider a ball rolling in a friction-less U-bend. When dropped from one side, the ball rises to the same height on the second. As the second side is stretched out the ball has to travel a greater distance to re-obtain the height from which it fell. In the limit, if we flatten the second side, the ball will have to travel an infinite distance in an attempt to regain its height.

The first point to note is that thought experiments that involve idealisation often resemble the limiting case of the extrapolation of experimental results. When trying to prove a general law, scientists often plot the results of some performed experiments on a graph and then interpolate and extrapolate from these. Intuitively, the infinite number of possible experiments represented by the points on the line are not thought experiments, although the limiting case of a series of experiments performed with ever decreasing amounts of friction, say, may well be a thought experiment. The difference, it seems to me, lies in the fact that in the case of the thought experiment the experimenter visualises or describes some hypothetical situation, whereas the extrapolator does not imagine the infinite number of possible experiments that concern him. However, aside from the visualisation element, extrapolation and thought experiments involving idealisation are very similar. Critics of thought experiments that involve idealisation are going to have a

tough time saying why extrapolation is justified (as they must, if they are not to reject much science) but such thought experimentation is not.

Interestingly, thought experiments that aim at exploring our concepts and values by describing some situation and then asking us what we would say or do involve idealisation. Imagine a situation in which an agent is faced with having all her fingernails pulled out one by one by a sadistic but powerful torturer. She can stop the torture at any point by pressing a red button that will trigger a nuclear explosion and thus kill everyone, herself included. Often when such thought experiments are put to us we are asked what we would do. However, I suggest, what we would do is not really the issue. In several (maybe most) near-by possible worlds in which I have my fingernails pulled out, I loose all self-control under the pain and press the button. In some of these possible worlds, you are just as pathetic. Thus, rather than it being relevant what we would say or do, we should be interested in what some ideally calm, good, and rational person would say or do. In such cases thought experiments trump real experiments. The judgements of people contemplating what should be done under torture are more reliable than the judgements of people actually being tortured.

Other thought experiments cannot be replaced by real experiments because they involve the violation of physical laws. Thus they cannot be performed, nor even approximated. The purpose of such thought experiments is to shed light on logical possibility. Such thought experiments are very similar to the computer simulations that scientists often run in order to see how events would unfold if the laws of nature were slightly different.[8] For example, physicists can use computers to model how the universe would unfold if G, the gravitational constant, were different. The thought experimenter models a different world in her head; the simulator uses a computer. At

least in simple cases, a thought experimenter with a clear grasp of the relevant laws should be as reliable as a computer simulation. Philosophers who are suspicious of thought experiments in which physical laws are violated are going to need to provide a reason why models produced by people are worthless, but computer simulations can be trusted (and they must trust computer simulations, or once again they are forced to reject much scientific practice). I suggest no such reasons will be forthcoming, and that simulations and physically impossible thought experiments should both be considered potential sources of knowledge.

## Conclusion

In this paper I have presented an account of thought experiments. According to this account, a thought experimenter manipulates her world view in accord with the "what if" questions posed by a thought experiment. When all necessary manipulations are carried through the result is either a consistent model, or contradiction. If a consistent model is achieved the thought experimenter can conclude that the scenario is possible, if a consistent model cannot be constructed then the scenario is not possible.

The account differs from Nersessian's and Miščević's model-based accounts in various ways. Most importantly, their accounts claims that thought experimenters use mental models, of a type posited by some psychologists. In contrast, in my account the nature of the model used by the thought experimenter is unconstrained.

I have suggested that thought experiments can fail in two ways: The thought experimenter may be unable to answer the "what if" questions, or the thought experimenter may make a mistake as to whether she has constructed a consistent or

inconsistent model. Despite their fallibility, however, thought experiments can enable us to gain knowledge. Those who argue that they should be replaced by real experiments are mistaken.

**Acknowledgements**

Thanks are due to Peter Lipton and Joel Katzav, who read and commented on drafts of this paper. Versions of this paper have been presented at Birmingham University and at the Joint Session of Mind and the Aristotelian Society. I have benefited greatly from the comments of those present.

**Mailing address for author**

IEPPP

Furness College

Lancaster University

Bailrigg

Lancaster

LA1 4YW

U.K.

**E-mail:** R.V.Cooper@lancaster.ac.uk

**References**

Benacerraf, P. 1973. "Mathematical Truth." Reprinted in *The Philosophy of Mathematics* (1996), edited by W.D.Hart, 14-30. Oxford: Oxford University Press.

Bishop, M. 1999. "Why Thought Experiments are not Arguments." *Philosophy of Science* 66, 534-541.

Brown, J. 1991a. *The Laboratory of the Mind.* London: Routledge.

———. 1991b. "Thought Experiments: A Platonic Account." In *Thought Experiments in Science and Philosophy,* edited by T. Horowitz and G. Massey, 119-128. Savage: Rowman and Littlefield Publishers.

Bunzl, M. 1996. "The Logic of Thought Experiments." *Synthese* 106, 227-240.

Davenport, E. 1983. "Literature as Thought Experiment (On Aiding and Abetting the Muse)." *Philosophy of the Social Sciences* 13, 279-306.

Galileo, G. 1974. *Two New Sciences.* Translated by S.Drake. Madison: University of Wisconsin Press.

Gendler, T. 1998. “Galileo and the Indispensability of Scientific Thought Experiment.” *British Journal for the Philosophy of Science* 49, 397-424.

Gooding, D. 1990. *Experiment and the Making of Meaning.* Dordrecht: Kluwer.

Gordan, R. 1986. “Folk Psychology as Simulation.” *Mind and Language* 1, 158-171.

Häggqvist, S. 1996 *Thought Experiments in Philosophy.* Stockholm: Almqvist and Wiksell International.

Hull, D. 1998. “That Just Don’t Sound Right.” In *The Cosmos of Science: Essays of Exploration,* edited by J. Earman and J. Norton, 430-457. Pittsburgh: University of Pittsburgh Press.

Hume, D. 1978 [1888]. *A Treatise of Human Nature.* Oxford: Clarendon Press.

Humphreys, P. 1993. “Seven Theses on Thought Experiments.” In *Philosophical Problems of the Internal and External Worlds,* edited by J. Earman, 205-227. Pittsburgh: Pittsburgh University Press.

John, E. 1998. “Reading Fiction and Conceptual Knowledge: Philosophical Thought in Literary Context.” *Journal of Aesthetics and Art Criticism* 56, 331-348.

Kuhn, T. 1964. “A Function for Thought Experiments.” Reprinted in *Scientific Revolutions* (1981), edited by I. Hacking, 6-27. Oxford: Oxford University Press.

Mach, E. 1960. *The Science of Mechanics.* La Salle: Open Court.

Maffie, J. 1997. “‘Just-So’ Stories About ‘Inner Cognitive Africa’: Some Doubts About Sorensen’s Evolutionary Epistemology of Thought Experiments.” *Biology and Philosophy* 12, 207-224.

Maudlin, T. 1994. *Quantum Non-locality and Relativity.* Oxford: Blackwell.

McAllister, J. 1996. “The Evidential Significance of Thought Experiment in Science.” *Studies in History and Philosophy of Science* 27, 233-250.

Miščević, N. 1992. “Mental Models and Thought Experiments.” *International Studies in the Philosophy of Science* 6, 215-226.

Nagel, T. 1998. “Conceiving the Impossible and the Mind-Body Problem.” *Philosophy* 73, 337-352.

Nersessian, N. 1992. “In the Theoretician’s Laboratory: Thought Experiments as Mental Modelling”, in *Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association. Vol. 2.* (1993), edited by D.Hull, M.Forbes and K.Okruhlik, 291-301. Michigan: Philosophy of Science Association.

Norton, J. 1991. “Thought Experiments in Einstein’s Work.” In *Thought Experiments in Science and Philosophy,* edited by T. Horowitz and G. Massey, 129-148. Savage: Rowman and Littlefield Publishers.

———. 1996. “Are Thought Experiments Just What You Thought?” *Canadian Journal of Philosophy* 26, 333-366.

Poincaré, H. 1952. *Science and Hypothesis*, translated by W. Greenstreet. New York: Dover Publications.

Sidelle, A. 1998. “Review of Häggqvist’s Thought Experiments in Philosophy” *The Philosophical Review* 107, 480-483.

Sorensen, R. 1992a.*Thought Experiments*. Oxford: Oxford University Press.

———. 1992b. “Thought Experiments and the Epistemology of Laws” *Canadian Journal of Philosophy* 22, 15-44.

Williams, B. 1973. *Problems of the Self*. Cambridge: Cambridge University Press.

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

**Notes**

[1] Brown, Gooding, Kuhn, Mach, McAllister, Nersessian and Norton concern themselves with thought experiments in science; Sorensen and Wilkes write about thought experiments in all areas (although Wilkes is sceptical about the value of thought experiments in philosophy).

[2] Bunzl (1996) claims that all knowledge producing thought experiments are deductive arguments. I reject his account for the same reasons that I reject Norton's. Häggqvist (1996) presents an account whereby thought experiments are not arguments but "work only through their connection with arguments". Sidelle (1998) convincingly argues that when the details are spelt out Häggqvist's view collapses into the claim that thought experiments are arguments.

[3] Other writers have given other reasons for rejecting Norton's account. Gendler (1998) shows that one of Galileo's thought experiments cannot be construed as an argument. Bishop (1999) argues that Norton's account cannot account for cases where people disagree about the results of a thought experiment. In such cases the parties reconstruct the thought experiment as two different arguments, but they are discussing the same thought experiment.

[4] Brown provides a brief overview of his account in Brown (1991b.)

[5] Brown argues against a causal account of knowledge on the basis of the EPR experiment. He claims that we gain knowledge about the electron's mate but there is no causal link between the two electrons. However, theories involving tachyonic connections between the electrons would supply the missing causal link (Maudlin 1994).

[6] I accept that pieces of fiction may count as thought experiments, so long as the "what if" questions are rigorously followed through. See Davenport (1983) and John (1998) for discussion of literature as thought experiment.

[7] Writers who have argued that real examples should be used instead of thought experiments include Hull 1997. Wilkes 1988 is sceptical of thought experiments in philosophy.

[8] Humphreys 1993 p.219 also notes that such computer simulations are like thought experiments.