Studies in History and Philosophy of Science 58 (2016) 24–33

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Taming theory with thought experiments: Understanding and scientific progress

Michael T. Stuart

Center for Philosophy of Science, University of Pittsburgh, Cathedral of Learning, Room 817Q, Pittsburgh, PA, 15260, USA

ARTICLE INFO

Article history:
Received 20 November 2015
Received in revised form
7 April 2016

Keywords:
Scientific understanding;
Thought experiments;
Scientific progress;
Maxwell's demon;
Einstein's clock in the box;
Darwin's vertebrate eye

ABSTRACT

I claim that one way thought experiments contribute to scientific progress is by increasing scientific understanding. Understanding does not have a currently accepted characterization in the philosophical literature, but I argue that we already have ways to test for it. For instance, current pedagogical practice often requires that students demonstrate being in either or both of the following two states: 1) Having grasped the meaning of some relevant theory, concept, law or model, 2) Being able to apply that theory, concept, law or model fruitfully to new instances. Three thought experiments are presented which have been important historically in helping us pass these tests, and two others that cause us to fail. Then I use this operationalization of understanding to clarify the relationships between scientific thought experiments, the understanding they produce, and the progress they enable. I conclude that while no specific instance of understanding (thus conceived) is necessary for scientific progress, understanding in general is.

One only understands the things that one tames.

—A fox (Saint-Exupéry, 1943)

I claim that one way thought experiments enable scientific progress is by increasing understanding. To make this claim, I need to say something about what understanding is. Despite philosophical interest,[1] there is no currently accepted characterization of understanding. The most general characterization defines understanding as any epistemologically desirable state that is not knowledge. But this needs to be more specific if it is to do any philosophical work. Some philosophers characterize understanding as whatever a good explanation provides (e.g., Salmon, 1984; but see Lipton, 2009 for counterexamples). Others as what we get when we reduce the number of fundamental entities that we have to admit in a theory (e.g., Friedman, 1974), or what happens when we find a way to explain different phenomena using the same patterns of argument (Kitcher, 1981). Henk de Regt characterizes understanding in terms of intelligibility, which is "the value that scientists attribute to the cluster of virtues (of a theory in one or more of its representations) that facilitate the use of the theory for the construction of models" (2009, p. 31). Hasok Chang claims that understanding "is knowing how to perform an epistemic activity" (Chang 2009, p. 75). This definition brings us to the ongoing debate concerning whether understanding is a type of knowledge. For example, Lipton (2004) argues that understanding is knowledge of causes, and Grimm (2006) argues that it can be Gettiered. But while many agree that whatever understanding turns out to be, it will be a kind of knowledge (Achinstein, 1983, p. 23; Kitcher, 2002; Salmon, 1989, pp. 134-5; Woodward, 2003, p. 179), others are not so sure (Elgin, 1996, 2004; Kvanvig, 2003; Zagzebski, 2001).

Deciding between these definitions isn't necessary for my purposes. All I need is a way of picking out intuitive instances of understanding to show that thought experiments can provide

E-mail address: m.stuart@pitt.edu.

[1] E.g., Achinstein (1983), De Regt (2009), De Regt and Dieks (2005), De Regt, Leonelli, and Eigner (2009), Elgin (1993, 2004), Friedman (1974), Kitcher (1981), Kosso (2006), Machamer and Woody (1994), Salmon (1984), Toulmin (1972), van Fraassen (1980), and Woodward (2003).

http://dx.doi.org/10.1016/j.shpsa.2016.04.002
0039-3681/

something like what we get in those instances. I will pick out instances of understanding by using the following consideration: we already can, and do, *test* for understanding, both in ourselves and in our students. There are at least two kinds of test that we use to do this.

1. Two tests of understanding

The following two tests can be identified in many fields of study.

In the first, we are asked to demonstrate a grasp of the meaning of a concept, idea, or theory. This type of test may come in the form of true/false questions, matching questions, multiple choice questions, definition questions, short answer questions, or (the first half of) an essay. Consider a few examples:

1) True or false? A valid argument must have all true premises.
2) Which of the following is *not* one of Aristotle's four causes?
3) What is constructive empiricism?

In order to answer these questions correctly, we need to have semantic relations established between the ideas in question and our existing ideas, concepts, and experiences. For the first question, we have to connect the concept VALID with some logical definition, and ask ourselves what that definition says about true premises, which itself requires that we know what premises are, and so on. Establishing such relations is one of the central aims of education.

My first claim is that the epistemic state that enables us to answer questions of this type is understanding. If we were looking for knowledge, we might ask for something like a list of justified true beliefs. Instead, this type of question asks for evidence of semantic digestion. If we could *only* repeat the textbook definition of some concept, theory or model, we can't be said to understand it, since in this case, we could only answer questions that require exact repetition of that definition. Such repetition would count as a display of knowledge, but not understanding, and the semantic question type usually asks for more than this. True/false, multiple choice, and compare and contrast questions require the ability to make distinctions and draw relations within and between concepts. If I can't say how constructive empiricism differs from realism or positivism, I probably don't understand constructive empiricism.

The strongest evidence for this sort of understanding is the ability to put a theory, concept or model in our own words, and explain it to others. If I can do this, then at some point I must have interpreted the new idea with respect to my existing ideas and experience. Oral examinations and job interviews require so much preparation because they are among the most fine-grained tests of this sort of understanding.

In any case, honest success with *any* of the above question types requires the existence of meaningfully formed relationships between new ideas and existing ones. Let's call the set of tests that primarily rely on the establishment of semantic relations, tests of *meaningfulness*.

We encounter a second kind of test for understanding when we are asked to *do something* with an idea, concept, theory or model. This kind of test might require that we argue for a conclusion, derive a result, play a piece of music, or disassemble a handgun. Passing this sort of test requires that the new idea have found its way into our cognitive toolkit. In other words, we must be able to achieve something that we could not have achieved before, or could not have achieved as efficiently without using the new idea. For example, I may have been able to provide moral arguments for abortion, but I might not have been able to do so explicitly using utilitarian reasoning. And this sort of ability is partially what is required to say that I understand utilitarianism. Call the set of tests that require us to demonstrate a new ability, tests of *fruitfulness*.

For full marks, most written tests require that we display both sorts of understanding. First we show that we understand the new ideas in terms of the relations between them and our previous ideas and experience. Then we are asked to do something with them.

Finally, the fact that we have a range of grades for success in these two endeavours reflects the fact that there are grades of understanding. The more deeply we've sewn a new idea into our doxastic quilt, the better. And the more problem types to which we can apply the new idea, the better.

A few caveats. It is still possible to pass both types of test without having any real understanding. Tests are always imperfect, and there are as many ways to fake understanding as there are knowledge. I introduce these tests, therefore, merely as a way of operationalizing understanding: we have the experience of passing these two types of test, we know what it feels like to transition from encountering a new term for the first time, coming to grasp its meaning (shallowly then deeply), and learning to use the new term, theory or model to do something.

Second, there will be genuine instances of understanding that this operationalization does not capture. However, all I need for my argument is that if *S* has the ability to pass these two tests (to some minimally high degree) with respect to *p*, then we have reason to think that *S* understands *p* (to some minimally high degree).

Third, there may be relations between the cognitive states tested by each of the two tests. For example, it might be the case that to pass the fruitfulness test, we must also be able to pass the meaningfulness test. I do not want to make any claims about such relations at this time. Finally, while I am interested in the cognitive mechanisms required to pass these tests, I make no claim about them here.[2]

My main argument in this paper is the following. Some thought experiments enable understanding in that they help us to pass the meaningfulness test and the fruitfulness test. Being able to pass these tests is necessary for scientific progress. Therefore, some thought experiments can enable scientific progress by increasing understanding. The first premise is supported by case studies. The second by a short argument along the following lines: it is prima facie plausible to think that we cannot make progress with a new scientific idea if we do not know what it means and cannot achieve anything with it.

I should say that the arguments in this paper are orthogonal to most of the existing literature on thought experiments. It has generally been granted since Kuhn (1977, p. 263) that thought experiments contribute to scientific progress, although there are skeptics.[3] Scientific progress can be understood as the accumulation of new propositional knowledge (as in Bird, 2007, 2008), and some philosophers (including Brown, 2004, p. 34; Gendler, 2004, p. 1152; Kuhn, 1977, p. 241; Norton, 2004, p. 44; Thagard, 2010, p. 251) have discussed the way that thought experiments might make this possible. Another way to characterize scientific progress is as an increase in understanding (as in Bangu, 2015; Dellsén, 2016), and some philosophers have discussed the way that thought experiments increase understanding (e.g., Arthur, 1999; Camilleri, 2014; Gendler, 1998, 2000; Gooding, 1993, 1994; Humphreys, 1993; Lipton, 2009; Nersessian, 1992, 2007). None of these, however,

[2] In Stuart (forthcoming) I argue that one thing necessary for passing both tests is imagination.

[3] These include Meinong (1907), Duhem (1954, pp. 201-205), Dancy (1985), Harman (1986), Thagard (2010, 2014) and Wilkes (1988). For replies, see e.g., Häggqvist (1996, chap. 2) and Stuart (2014).

provide characterizations of understanding, nor are their arguments carried further to considerations of scientific progress. Nor are the cases that I will present considered in the literature in terms of understanding (with the exception of Maxwell's demon by Brown, 1993, p. 274, and Heisenberg's microscope by Camilleri, who is cited below). For the purposes of this paper, I do not deny that thought experiments can provide knowledge, nor do I deny any of the accounts in the literature created to explain how they do. I also do not make any claims about the nature of thought experiments themselves, for example, whether they are arguments or mental models. I do claim, however, that my operationalization of understanding is plausible for explaining how scientists generate new understanding in science using thought experiments, and I claim that it helps us appreciate how some very important thought experiments contributed to scientific progress.

I'll now present three cases that have been historically instrumental in passing the tests of meaningfulness and fruitfulness, and two failures.

2. Case studies

2.1. Maxwell's (original) demon: pass

Maxwell's demon does not appear with Maxwell's now-famous papers on statistical thermodynamics (Maxwell, 1860a, 1860b, 1860c, 1860d, 1861, 1866). Rather, the demon first appears in an 1867 letter from Maxwell to Peter Guthrie Tait, and then again in a letter to J. W. Strutt in 1869. The thought experiment containing the demon is then published in *A Theory of Heat*, in 1871, a book whose aim is "to exhibit the scientific connexion of the various steps by which our knowledge of the phenomena of heat has been extended" (Maxwell, 1871, p. v). In other words, this book presumes that statistical thermodynamics is well-supported both theoretically and empirically, and is now ready to be communicated to a more popular audience.

It seems unlikely therefore that Maxwell was using the demon as an argument for the truth of his statistical theory of heat, especially since the demon does not appear until the end of the final chapter, three pages from the end of *A Theory of Heat*'s more than three hundred and forty. Maxwell says at the beginning of this final chapter that "We have already shown that heat is a form of energy that when a body is hot it possesses a store of energy, part at least of which can afterwards be exhibited in the form of visible work" (1871, p. 308). He considers his theory of heat to have been demonstrated by this stage of the discourse, and now he will ask interpretive questions. For instance, Maxwell argues that we should portray heat energy as kinetic energy. But kinetic energy of what? To answer, Maxwell turns to the molecular theory of matter. Molecules are characterized as the smallest part of something that retains the properties of that thing (Maxwell, 1871, p. 313). These will be the carriers of kinetic energy.

However, we must distinguish between the kinetic energy of a molecule and the average kinetic energy of a mass of particles. Failing to make this distinction could lead us to misunderstand his theory. In Maxwell's words, "it is therefore possible that we may arrive at results which, though they fairly represent the facts as long as we are supposed to deal with a gas in mass, would cease to be applicable if our faculties and instruments were so sharpened that we could detect and lay hold of each molecule and trace it through all its course" (pp. 315-16).

Maxwell employs two examples to make this distinction between levels of analysis as clear as possible. First, he asks us to compare the overall impact of education on society with the impact of education on a single student. There may be no student whose experience reflects the average effect of education on her or his society, but this does not affect the average impact of education on society. Both the average and individual effects of education are sensible objects of study. Second, we are asked to compare the average accuracy of a firing squad with the path of a single bullet (p. 316). Again, perhaps no single bullet possesses an accuracy equal to the accuracy of the firing squad, but this conflict does not invalidate either measurement. Maxwell is appealing to the imagination of his readers to help them understand how the statistical theory of heat can apply to experience as a statistical and aggregative measure, and to reiterate that it need not compete with everyday experience.

To further support this claim, Maxwell derives experimental generalities concerning gases from the new theory of heat, including Boyle's law, Guy-Lussac's law, Charles's law and Dulong and Petit's law (pp. 321-333). Combined with the above imaginary examples, Maxwell forges a powerful two-way connection between theory and experience. First, our everyday experience with bullets and students helps us understand how his theory relates to experience. Second, the theory explains regularities (like Boyle's law, etc.) that experience has taught us.

The demon arrives at last. "Before [he] conclude[s]," Maxwell "wants to draw attention to one more aspect of the molecular theory which deserves consideration":

> One of the best established facts in thermodynamics is that it is impossible in a system enclosed in an envelope which permits neither change of volume nor passage of heat, and in which both the temperature and the pressure are everywhere the same, to produce any inequality of temperature or of pressure without the expenditure of work. This is the second law of thermodynamics, and it is undoubtedly true as long as we can deal with bodies only in mass, and have no power of perceiving or handling the separate molecules of which they are made up. But if we conceive a being whose faculties are so sharpened that he can follow every molecule in its course, such a being, whose attributes are still as essentially finite as our own, would be able to do what is at present impossible to us. For we have seen that the molecules in a vessel full of air at uniform temperature are moving with velocities by no means uniform, though the mean velocity of any great number of them, arbitrarily selected, is almost exactly uniform. Now let us suppose that such a vessel is divided into two portions, A and B, by a division in which there is a small hole, and that a being, who can see the individual molecules, opens and closes this hole, so as to allow only the swifter molecules to pass from A to B, and only the slower ones to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics. (1871, pp. 338-339).

This is the thought experiment, which clearly forms part of the same chapter-long project. To be sure we appreciate this, Maxwell summarizes this project once again, immediately following the thought experiment:

> This is only one of the instances in which conclusions which we have drawn from our experience of bodies consisting of an immense number of molecules may be found not to be applicable to the more delicate observations and experiments which we may suppose made by one who can perceive and handle the individual molecules which we deal with only in large masses. (1871, p. 339)

Like the derivations which produced the experimental regularities from the molecular theory, the thought experiment deals with the relation between the kinetic theory of heat and our expected experience of an empirical regularity known to obtain; in this case, the second law of thermodynamics. However, Maxwell will not (or cannot) straightforwardly derive the second law from the kinetic theory of heat. In fact, Maxwell *never* presented a mathematical treatment of the second law, at least not in any of his written works or correspondence (Garber, Brush, & Everitt, 1996, p. 59).

Here is Maxwell's position: He knows that statistical thermodynamics is empirically adequate and explanatorily powerful. And he knows that it allows violations of the second law. He also knows that we cannot create a perpetual heat engine, which we should be able to if the second law were violable (since we could increase the heat of a substance without adding heat, and use that heat to run an engine). Maxwell must therefore provide some explanation of how the second law considered as merely statistically valid connects to our inability to produce a perpetual heat engine. But *unlike* the intuitive examples or the theoretical derivations of laws, this connection between theory and empirical regularity cannot proceed either by citing everyday experience or by mathematical reasoning. He cannot explain what the second law of thermodynamics considered statistically means using a case from everyday experience because we have no experience with very large numbers of things all with the same mass but varying velocities that encounter each other through forces that can act at a distance. And we cannot (even to this day) derive the second law of thermodynamics from something more fundamental. Maxwell needs something else to explain what sort of world the statistical theory of heat implies. And he chooses a demon to assist the imagination in making that leap from what we take ourselves to know (the second law and the adequacy of the statistical theory of heat), to something we *understand*.

The understanding gained comes from exploring in our imagination the situation Maxwell provides. We envisage a tiny being that can see and interact with individual molecules directly. We find no reason why such a being could not purposely isolate faster moving molecules from slower, and thereby increase the heat of a region for free, because if *we* were that being, then there would be no reason *we* could not. We may have trouble imagining a being that can see molecules, but if we imagine *ourselves* in an analogous position, say, in control of a sliding door, surrounded by molecules which act like medium sized rubber balls, we understand the scenario perfectly. And a being in such a scenario could increase the average kinetic energy of one partition of a container and thereby create a perpetual motion machine. At the same time, we appreciate why such a scenario, which represents *all* violations of the second law, will not arise in practice. This is because as far as we know, there is nothing that can discern the positions of molecules without also adding energy to that system, and perhaps there could not be. While random increases in entropy should be expected at tiny intervals across tiny spaces, the odds of such occurrences taking place in a way that is noticeable, regular, and useful for humans, is so low as to be negligible. This recovers our everyday inability to produce perpetual heat engines, and also clarifies the second law by interpreting it according to Maxwell's kinetic theory of heat. As Wayne Myrvold argues (2011), the demon allows Maxwell to deny the validity of the second law traditionally conceived, but uphold the *practical* inability to do what the demon does.

Now let's see how this thought experiments helps us pass the meaningfulness and fruitfulness tests.

Through the thought experiment, we establish meaningful semantic relations between our concepts in a way that explains why the second law of thermodynamics considered statistically does not violate our everyday inability to produce perpetual heat engines. We understand that only something like the demon can create the kind of violation that would be needed in order to profit mechanically from the statistical nature of the second law, and we understand this because we accept the demon as a representation of all violations of the second law, and we can put ourselves in its shoes. This creates the necessary semantic ties. Indeed, the demon is still used in textbooks and on educational websites precisely because it helps us relate a merely statistical second law to notions and experiences that we already have. Thanks to the demon, we make these otherwise difficult semantic relations despite any cognitive difficulties we might have with molecules or statistics.

Second, we gain several important abilities concerning the second law considered statistically. We can now *explain* why we cannot build perpetual heat engines despite statistical violability of the second law. And we can ask new experimental and theoretical questions: could something like a demon actually exist? What properties would such a demon have? This is precisely the direction that the discussion following Maxwell took (see, e.g., Norton and Earman, 1999a, 1999b). In other words, thanks to the demon, we can offer new explanations, we can ask new theoretical questions about heat, and we can perform new experiments that might not otherwise have occurred to us.

2.2. The clock in the box: pass

The clock in the box thought experiment of Albert Einstein's presented at the 1930 Solvay conference tries to show the incoherence of quantum theory by targeting the uncertainty principle proposed by Werner Heisenberg in 1927 (see Kragh, 2002, pp. 212-213). Einstein apparently presented the clock in the box "to evade the fourth Heisenberg uncertainty relation" (Treder, 1975, p. 135), which states that the product of the "uncertainty in the knowledge" (Heisenberg, 1930, p. 16) of the energy for a particle with the uncertainty in the knowledge of the time of measurement of that particle will be no less than the reduced Plank constant.

As in Maxwell's context, we have a newly proposed theory that is gaining theoretical and empirical support. By 1930, even Einstein had accepted the position-momentum uncertainty principle as true. And this case is even stronger than Maxwell's, since the uncertainty relations were theoretically *derivable*. In other words, there was then (as now), little doubt concerning the empirical adequacy or theoretical grounding of the uncertainty relations. We might say we know them.

What we don't know is what sort of world they imply, or their modal status. Is uncertainty a product of something more fundamental, which isn't itself statistical? Is there something empirical that unites the different uncertainty principles? I claim that Einstein's thought experiment is created to address questions like these concerning the meaning of the uncertainty principles.

Here is Einstein's thought experiment as presented by Niels Bohr (1949). Suppose we have a box containing a source of radiation. There is an aperture in the side of the box, covered by a shutter which is controlled by a clock. The clock opens the shutter for an arbitrarily precise amount of time, just enough to let a single photon be emitted (see Fig. 1). Since the clock controls the shutter, we can specify the exact time of emission. Now suppose that we weigh the box before and after the photon is emitted. This too can be done with arbitrary precision. Since $E = mc^2$, we can determine the energy of the system before and after the emission—that is, with and without the photon (by extracting a mass measurement from the weight measurement). We subtract the two energies (before and after), and get the energy of the photon. Since we know the exact time of emission, we obtain a contradiction with the uncertainty principle.

The thought experiment was "quite a shock to Bohr," nevertheless with the "next morning came Bohr's triumph" (Rosenfeld, quoted in Pais, 1982, pp. 446-447). Different authors present Bohr's triumph in different ways, but it is agreed that the first step was to

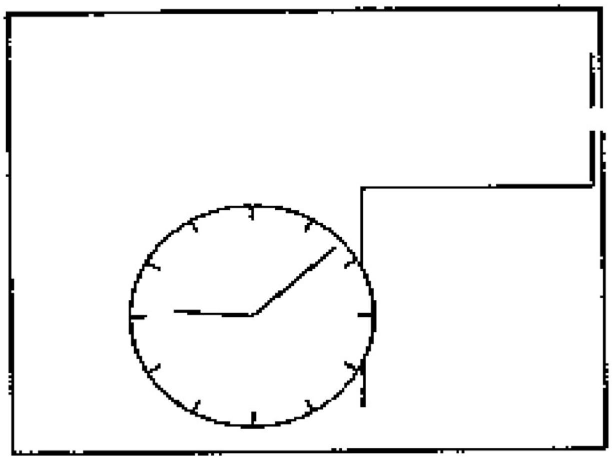

Fig. 1. Einstein's clock in the box, depicted by Bohr (Bohr, 1949, p. 225).

provide a detailed analysis of the experimental set-up. To this end, Bohr modified Einstein's scenario. A "pseudo-realistic" drawing (Bohr, 1949, p. 226) helped Bohr demarcate the crucial differences between Einstein's set-up (Fig. 1) and his own (Fig. 2).

Bohr wanted to look closely at the measurement process that would be required to determine the mass of the emitted photon. This might be surprising, because that aspect of the thought experiment seems the least problematic: we simply weigh the box. But it makes sense if we consider Bohr's general philosophy of quantum mechanics: measurement requires action on the part of measuring agents, whose influence on the system is inseparable from the system. "The problem again emphasizes the necessity of considering the *whole* experimental arrangement, the specification of which is imperative for any well-defined application of the quantum-mechanical formalism" (Bohr, 1949, 230). In other words, Bohr's justification depends on presenting the scenario in what he thinks is a more complete and detailed way.

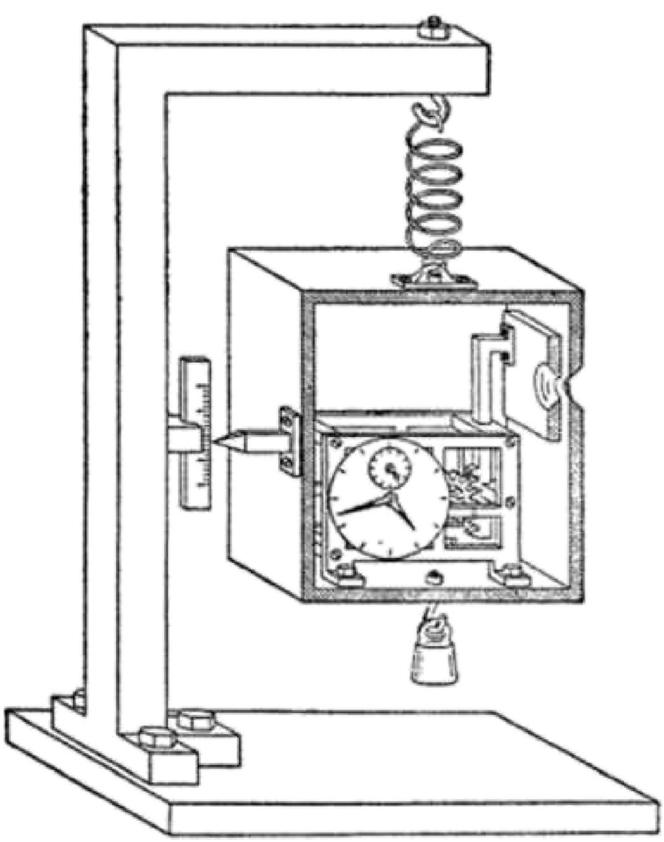

Fig. 2. Bohr's pseudo-realistic drawing of the clock in the box (Bohr, 1949, p. 227).

Following the experimental set-up shown in Fig. 2, we measure the initial weight of the box by the position of the pointer on the left of the box. We then release the photon, and find its weight by adding weights to the bottom of the box, until we find one weight that returns the box exactly to its original position on the scale. Weighing the box in this way requires the presence of a gravitational field (to pull the box down on the spring). And due to Einstein's own work in relativity, the clock inside the box will slow down or speed up depending on its speed and direction of motion in the gravitational field. Bohr's derivation suggests that the more precise our measurements of the momentum of the box, the less precise we can be about our time measurements. And therefore we cannot violate the uncertainty principle this way because the uncertainty creeps in through the movement of the clock in a gravitational field (for a reconstruction of the formal argument, see Bishop, 1999).

Presented with this reply, Einstein permanently "accepted fully all the Heisenberg relations" (de la Torre, Daleo, & Garcia-Mata, 2000, p. 54), and never brought up the matter again. However, like Maxwell's demon, the clock in the box will not die; it continues to be debated in the same context as it was originally presented, namely, concerning whether and how the uncertainty principle maintains itself in the face of certain possible experimental arrangements (see Hilgevoord, 1998; Hnizdo, 2002; Kudaka & Matsumoto, 1999; Treder, 1970).

And this continuing discussion is especially interesting given that the uncertainty principle can be, and was, even in 1930, derived theoretically and confirmed experimentally. Again, like Maxwell's demon, this thought experiment operates first by helping us to understand not just *that* the uncertainty principle is safe from Einstein's set-up, but also why it will not be violated by any such set-up. And in doing this, it tells us something about the sort of world described by the new quantum theory.

The clock in the box thought experiment can be seen as an attempt to connect the uncertainty principle either to experience or to parts of physical theory. Bohr connects the idea motivating the thought experiment to an imagined physical set-up that we can look at, but also to parts of relativity theory. Each modern instance of the experiment establishes its own connection between the time–energy uncertainty principle and some (set of) clocks in boxes. Making these connections helps to expose the meaning of the uncertainty principle and the nature of its relation to experience. It has been almost a century since it was introduced, and physicists still do not agree whether the uncertainty principles set a maximum level of precision on measurements of pairs of conjugate variables, knowledge of those quantities, the definition of those variables, or the statistical spread of those variables (Hilgevoord, 2006). The above-mentioned thought experiments aim at clarifying the meaning of one uncertainty principle by giving us non-theoretical, non-empirical reasons to believe that no matter what we do, uncertainty will always find a way back into the system.

"Here were two titans of modern physics with quite opposed positions, struggling to establish their view of the meaning of the quantum" (Norton, Unpublished, chap. 29). Like Norton, I want to interpret this clash between Einstein and Bohr as one concerning how we understand the *meaning* of the uncertainty principle and why it will not be violated. And this is an interpretation that emphasizes understanding in the sense of being able to pass the meaningfulness test.

The clock in the box also helps us to pass the fruitfulness test. We learn from considering the whole episode that measurement procedures must be considered in detail for any proposed violation of the uncertainty principle. If someone proposed a new way to violate one of the uncertainty relations, now we know what to do.

To summarize, the clock in the box is a series of thought experiments that can be characterized together as performing the same roles as Maxwell's demon: They seek to increase our understanding of the principle by helping us say what it means (when it holds and why), and learn how to justify and test it.

2.3. Darwin's vertebrate eye: pass

The *Origin of Species* was published in 1859, although we know Charles Darwin was thinking about how to present his theory of evolution by natural selection for a long time leading up to that. The eye thought experiment is expanded in each edition, but was present from the first. Darwin wanted people to understand his idea as clearly as possible since he rightly foresaw empirical, theoretical and theological backlash.

In Chapter 6 of the *Origin* ("Difficulties of the Theory"), Darwin introduces four potential problems. The second is, "Can we believe that natural selection could produce, on the one hand, an organ of trifling importance, such as the tail of a giraffe, which serves as a fly-flapper, and, on the other hand, an organ so wonderful as the eye?" (Darwin, 1872, p. 171). Darwin re-states this as the problem of "organs of extreme perfection and complication" (p. 186), and he focuses on the eye. He contends,

> To suppose that the eye with all its inimitable contrivances for adjusting the focus to different distances, for admitting different amounts of light, and for the correction of spherical and chromatic aberration, could have been formed by natural selection, seems, I freely confess, absurd in the highest degree. (p. 186)

Darwin isn't pretending to find his *theory* absurd, only its application to certain complex cases. It isn't clear how something as complex as the vertebrate eye could possibly be the result of such a slow, directionless process. If we evolved incrementally over millions of years, then at some point between now (with complete, functioning eyes) and the distant past (when there were no eyes) there must have been an intermediate stage. But eyes only work when they are complete, so the intermediate stages would not have been functional. And furthermore they would not have evolved unless specifically to produce a complete eye later on, which is impossible given that Darwin's evolution is directionless. Darwin sees only one way to address this issue: "if numerous gradations from a perfect and complex eye to one very imperfect and simple, each grade being useful to its possessor, can be shown to exist… then the difficulty of believing that a perfect and complex eye could be formed by natural selection, though insuperable by our imagination, can hardly be considered real" (pp. 187-188).

To find these gradations, Darwin provides a thought experiment. We imagine a simple creature with no eyes. After a mutation, its offspring have a nerve somewhere on its body that has become sensitive to light. Such a nerve could help the creature discern how it was oriented in relation to the sun, or how far it was from the surface of the water, or whether a predator was looming, and so on. Any of these abilities would have been advantageous for this organism over its competitors. We imagine more nerves being added and subtracted randomly by additional chance mutations, over thousands of generations. A light-sensitive *patch* of nerves could register finer differentiations in light levels and would thus be more useful than individual nerves. We can also imagine topological changes to the patch and specialization of tissues. We can imagine the patch becoming concave. This would protect it from injury, without limiting its ability to gather light very much. As it became more and more concave, and therefore more protected, it might be filled with water or mucous. This would protect the light-sensitive cells, and slow down or focus the light. And at its most concave, it might almost form a complete recessed spheroid, at which point a lens would be advantageous. Muscles to focus the light and an increased number of nerves to process it could come later or concurrently. At each stage of this process, we have an organism with an adaptive mutation. The eye is no longer a counterexample.

The text of the actual thought experiment is less detailed, but the suggested thought experiment is clear (Darwin, 1872, pp. 187-188). Darwin writes,

> We ought in imagination to take a thick layer of transparent tissue, with a nerve sensitive to light beneath, and then suppose every part of this layer to be continually changing slowly in density, so as to separate into layers of different densities and thicknesses, placed at different distances from each other, and with the surfaces of each layer slowly changing in form. Further we must suppose that there is a power always intently watching each slight accidental alteration in the transparent layers; and carefully selecting each alteration which, under varied circumstances, may in any way, or in any degree, tend to produce a distincter image. We must suppose each new state of the instrument to be multiplied by the million; and each to be preserved till a better be produced, and then the old ones to be destroyed. In living bodies, variation will cause the slight alterations, generation will multiply them almost infinitely, and natural selection will pick out with unerring skill each improvement. (pp. 188-89)

The intentional language that Darwin uses would have been helpful for contemporary audiences, although today it is unnecessary. Darwin adds further credibility to the different stages of the thought experiment as he goes through them by pointing to the development of the human eye in the womb, which grows not completely but in stages, starting with very simple layers of translucent skin, and by comparing the eye to the telescope, which could be built up in stages from a single lens, where each stage involves a small improvement on the last, and at each we possess a functional telescope.

What is the purpose of this thought experiment? Darwin recognizes that "if it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down" (1859, p. 189). Darwin must therefore convince his readers that no such case will be found. He cannot provide any new theoretical argument for evolution that he has not already provided for the theory in general, and there is not enough fossil evidence to make his point empirically. Instead, Darwin picks a few exemplary cases, and shows us how to tackle all similar cases: by imagining steps that connect possible stages of development.

Just as Maxwell did with his demon, Darwin demonstrates a dual connection between theory and experience. Evolution by natural selection is explainable by reference to experience, and explanatory for experience. On the one hand, Darwin appeals to embryonic stages, the history of the telescope, and the eyes of simpler species to remind us what intermediary stages of ocular evolution might look like. On the other hand, there is no direct experience we can have of vertebrate ocular evolution, so Darwin invents a thought experiment that truncates the passage of time and millions of failed mutations. This enables us to grasp the relevant connections between our idea of evolution and our idea of organs of extreme perfection like the eye. The thought experiment also makes the theory more fruitful by showing us what to do when we are confronted by possible counterexamples, and gives us a way to begin our empirical investigations into evolutionary claims.

Like the demon and the clock in the box, Darwin's thought experiment presents both a problem and a solution. And like those cases, the solution comes not from a rigorous proof or empirical evidence, but from the *act* of imagining a hypothetical scenario. And, like the above cases, this is because Darwin is not in a position to tell us how the eye actually developed (as Maxwell could not derive the second law and Bohr could not prove that uncertainty is a fundamental element of reality and will always arise). Instead of new facts, Darwin gives us understanding of the conceptual machinery of his theory by giving examples that show us what evolution means and also how to use it. If we did not fully understand the idea of a stepwise accumulation of mutations, each of which face the battery of an economizing nature and so result in optimized traits and behaviours, we do by the end of Chapter 6 of the *Origin*. And if we want to know how to go about explaining other "organs of extreme perfection and complication," the method of Darwin's eye gives us a useful tool.

2.4. Heisenberg's microscope: fail

Heisenberg develops his matrix-mechanical formalization of quantum mechanics in (Heisenberg 1925). For this work, he is later awarded the Nobel Prize in physics "for the creation of quantum mechanics" (Nobel Prize in Physics, 1932). This paper was "the start of a new era in atomic physics, since any look at the physics literature of the next two or three years clearly shows the intensity and success of the work stimulated by this paper. His ideas were widely accepted" (Mott & Peierls, 1977, p. 221).

His microscope thought experiment appears in 1927, in truncated form, which Heisenberg later acknowledges as incomplete in an addendum to the paper. The paper is entitled "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik." Hilgevoord (2006) notes that "anschaulichen" has been translated in several ways, so that the title has become: "On the *Physical* Content of Quantum Theoretical Kinematics and Mechanics" (Wheeler & Zurek, 1983), "On the *Perceptible* Content of Quantum Theoretical Kinematics and Mechanics" (Blum, Dürr, & Rechenberg, 1984, listed in the references under Heisenberg, 1984), and "On the *Perceptual* Content of Quantum Theoretical Kinematics and Mechanics" (Cassidy, 1992). Hilgevoord himself considers "intelligible" or "intuitive" as substitutes. Searching for the physical, perceptible, perceptual, intelligible or intuitive content of quantum mechanics is quite in line with a characterization of Heisenberg's microscope as part of a project aimed at understanding.

The thought experiment is substantially re-worked for Heisenberg's lecture-tour of the United States in 1929. In these lectures, Heisenberg presents quantum mechanics to American scientists and university students. This was at a time when "quantum mechanics was…essentially complete, and the next task was to work out its consequences and to see how it would explain the many mysteries, paradoxes and contradictions in atomic physics" (Mott & Peierls, 1977, p. 226). The thought experiment is published in its completed form in 1930, in the book that resulted from these lectures, *The Physical Principles of Quantum Theory*.

Here it is. When we see an everyday object with our eyes, it is because we register visible wavelengths of electromagnetic radiation that bounce off it. The more precisely we want to determine the position of something by bouncing electromagnetic radiation off it, the shorter the wavelength of radiation we need. To determine the position of something very small, we need a very short wavelength. This is why for his thought experiment, Heisenberg imagines shooting gamma waves at a quantum particle to observe its position through a microscope: gamma waves have a very short wavelength. But the shorter the wavelength, the more energy is contained in the wave, which Heisenberg could treat mathematically as a particle thanks to his matrix-mechanical formalization of quantum mechanics. Instead of reflecting light off of a stationary object, we bombard a free particle with lots of high-energy particles in the form of gamma rays. By the time the electromagnetic radiation arrives at the lens of the microscope, the particle about which it carried its information will be long gone. Even if we knew the exact momentum of the incident gamma ray, we still have to deal with the Compton Effect, which tells us that the scattering of the particle after the collision creates an uncertainty that can only be reduced by selecting a type of electromagnetic wave with a longer wavelength. But in this case, we lose our precision in measuring the momentum of the particle. This is how Heisenberg uses an imaginary example to work through the position-momentum uncertainty principle. We cannot know the exact momentum of a particle at the same time as its position due to constraints placed on us by our strategies of measurement.

To understand the motivation behind the microscope, we require more historical context. Immediately after Heisenberg proved that his matrix-mechanical formulation and Schrödinger's wave-mechanical formulation of quantum mechanics were equivalent, Bohr wanted to unite quantum mechanics by producing an interpretation that was consistent with both wave and particle mechanics. Rather than supporting Bohr, however, Heisenberg sought to promote his own particle-focused interpretation. Since they were known to be mathematically equivalent, the only way to achieve this was, in my words, to increase either the meaningfulness or fruitfulness of the particle interpretation over the wave interpretation. "This [period] was a turning point for Heisenberg's theory, because it led him to propose a visualizable interpretation of quantum mechanics *through thought experiments* based on the limits of measurement. Heisenberg wrote out all his ideas in a letter to Pauli at the end of February [1927], in an attempt, *he said*, to 'get some sense of his own considerations' as he groped towards a consistent theory" (Beller, 1999, p. 105, my emphasis). Heisenberg's microscope therefore seeks a link between the new theoretical structure and some meaningful empirical content, for Heisenberg.

More specifically, Heisenberg's work left him with an equation that had p-values (physical values) and q-values (quantum values). He wanted to know what could be said about the q-values. What were they? Marten van Dyck writes:

> But then what does correspond in quantum mechanics to classical quantities like position? That is, *how are the q-numbers associated with physical quantities*, apart from their giving the right predictions about emitted spectra? The symbolic character of the new theory at first did not seem to allow an answer to these questions. This is why Schrödinger could refer to it as a 'formal theory of frightening, indeed repulsive, abstractness and lack of visualizability.' 'Heisenberg's theory in its present form is not capable of any physical interpretation at all,' was another claim made at the same time" (2003, p. 81).

This alarmed Heisenberg, and according to Van Dyck, was the reason he constructed the thought experiment. "The direct physical interpretation Heisenberg alludes to consists in the fact that the thought experiment allows him to see that the q-numbers need not keep their symbolic character, but can be given a conceptual content that is closely linked with their original kinematic meaning" (2003, pp. 81-85). In other words, the thought experiment gave Heisenberg a way to interpret the new theoretical formalism via a consideration of possible experiences that were visualizable or imaginable.

Here is one last piece of evidence concerning the role played by this thought experiment. Kristian Camilleri argues that

> Heisenberg's introduction of the imaginary gamma-ray microscope was not intended primarily to demonstrate the limits of precision in measurement. Though it certainly did this, its real purpose was to define the concept of position through an operational analysis. This becomes evident once we situate Heisenberg's use of the imaginary gamma-ray microscope within the context of his concerns over the meaning of concepts in quantum theory. (Kristian 2007, p. 179)

According to Camilleri, the microscope is meant to help us establish operational relations between "the concept of position" and experience. This thought experiment is therefore an excellent example of a thought experiment intended to help us pass the meaningfulness test.

The thought experiment *attempts* to produce this sort of understanding by connecting our everyday sort of knowledge to the theoretical structures of quantum mechanics. By imagining a microscope that interacts with a high-energy particle (that we can imagine as a round solid fast-moving object) that has recently collided with another moving particle, the thought experiment helps us see why we should not expect simultaneous measurement of position and momentum with arbitrary precision. However, it does this in a misleading way, and because of this, it fails. We cannot simultaneously measure the position and momentum of a quantum particle because of the intrinsically quantum nature of quantum systems, not because our apparatuses are not accurate enough to trace post-collision trajectories.

In sum, while the thought experiment does enable connections to be made between the principle and our other ideas and experience, they aren't the right connections. They aren't faithful to the causal structure of the system under consideration. And for the same reason, the new abilities that result from the thought experiment—to explain the empirical content or physical salience of the principle, or to seek out new applications or experimental tests—will lead us away from the truth if we exercise them. We gain new abilities, but they are not fruitful in the sense contextually established by scientists.

2.5. *Darwin's whale[4]: fail*

In a manuscript (and the first edition) of the *Origin*, Darwin presented the following thought experiment:

> In North America the black bear was seen by Hearne swimming for hours with widely open mouth, thus catching, like a whale, insects in the water. Even in so extreme a case as this, if the supply of insects were constant, and if better adapted competitors did not already exist in the country, I can see no difficulty in a race of bears being rendered, by natural selection, more and more aquatic in their structure and habits, with larger and larger mouths, till a creature was produced as monstrous as a whale (Darwin, 1859, p. 184).

In Louis Agassiz's copy of this manuscript,[5] "monstrous" is underlined twice, and the whole passage is marked with the words "This is truly monstrous!" Agassiz is right, we don't seem to gain any understanding from this thought experiment. And Darwin must have recognized this. Only a year later, in the second edition, the above passage simply reads: "In North America the black bear was seen by Hearne swimming for hours with widely open mouth, thus catching, almost like a whale, insects in the water" (1872, p. 184). The thought experiment has completely disappeared.

What is interesting about this failure is that the whale thought experiment is functionally equivalent to the eye: something complicated (whale behaviour and morphology) is made meaningful by reference to something we do understand (bear behaviour and morphology), plus a series of steps that can be imagined which lead us from the bear to the whale. But unlike the eye, it fails both tests. No additional meaningfulness is gained because we have no special difficulty imagining the ancestors of whales, and even if we did, imagining them to be bears is implausible. The thought experiment also fails the fruitfulness test. If we should try to explain the behaviour or morphology of highly complex living creatures by reference to existing living creatures, we would end up making needlessly strange evolutionary explanations. Being able to see bears as the ancestor of whales is just not biologically profitable.

Analysing these failures is important because they bring us to the relationship between scientific understanding and scientific progress.

3. Meaningfulness, fruitfulness and progress

Meaningfulness is necessary for progress because new ideas and experimental results do not tell us what they mean. No new idea will put itself into relation with our other ideas and experiences for us. We must always perform some interpretive action, however slight, whenever we are faced with a new piece of theory or empirical datum, whether as a student or as a scientist. Sometimes this action will be easy and automatic, and other times it will be difficult and require years of work and collaboration. I would argue that this action is an instance of solving what Bas van Fraassen calls the "problem of coordination," that is, coming to know what our theories, laws, models, concepts and equations are *about* (2008, p. 115). The problem of coordination may be solved trivially, as with Darwin's whale or Heisenberg's microscope, but these solutions will not count as genuine solutions for the scientific community. Instead, we require that the *right* sorts of relations between concepts and experience be made. Identifying the right relations is often a matter of theoretical, pragmatic, and sociological considerations, and trial and error. While the above arguments do not turn on how scientists identify the right connections, it does assume that they do. In my view, the real open question about the problem of coordination is not how it is solved at all, but how it is solved *well*.

Fruitfulness is necessary for progress because without the ability to use a new idea, we cannot begin to make new predictions, experiments and explanations. Again, this ability can be gained trivially by using a new idea in whichever way we like, or even the wrong way. I gain the ability to use CARBURETOR in a trivial sense if I decide to use it as a synonym for DECORATIVE OBJECT. This kind of behaviour may pass the fruitfulness test in the context of a game, but in the scientific context we are more careful about what counts as a successful use of a concept, theory, or model. And again, for this paper I do not require an account of how scientists distinguish between good and bad uses of a concept, I only require that they do, even if only implicitly, contextually and imperfectly.

Another reason to think that the understanding that enables us to pass the two tests is necessary for progress is that without it, we left in the dark, or worse, we go astray. In the case of Heisenberg's

[4] I am grateful to Andrew Inkpen for this example.

[5] Agassiz's marginalia retrieved from the Archives of the Museum of Comparative Zoology, Harvard University.

microscope and Darwin's whale, we begin to inquire in the wrong directions. We apply the wrong methods to the right systems, or the right methods to the wrong systems.

To summarize, without understanding, we are left with lists of empirical data that are disconnected from human aims and practices. And more data on its own will not remedy this. Scientific understanding, which I have characterized as related to the meaning and relevance of scientific data for scientists and society and the ability to apply that data broadly, is required for progress. This argument supports Finnur Dellsén's recent claim that scientific progress "matches" (Dellsén 2016, p. 72) or "follows" (Dellsén 2016, p. 82) increases in scientific understanding rather than knowledge. It also supports the claim that many philosophers have been making for some time now, that understanding should become the primary focus of epistemology (Dellsén, 2016; Elgin, 2006; Grimm, 2012; Kvanvig, 2003, 2009; Pritchard, 2009, 2010).

Whatever understanding is, it is clear that thought experiments can be helpful in gaining it. They can help us pass two tests of understanding when other methods fail. And the ability to pass these two tests is necessary in general for scientific progress, because without understanding thus conceived we cannot say what our theories or observations mean, and we cannot use them.

Acknowledgements

I'd like to thank Nancy Nersessian, John D. Norton, Yiftach Fehige, James R. Brown, Joseph Berkovitz, Marco Buzzoni, Catherine Elgin, Sören Häggqvist, Kenneth Westphal, Hans Radder, Martin Carrier, Elke Brendel, Nenad Miščević, Marcel Weber, Matthias Unterhuber, Maël Pégny, Agnes Bolinska, Andrew Inkpen, Leonardo Bich, Carol Cleland, Aaron Wright, Greg Lusk, Cory Lewis, and two anonymous referees for many helpful comments, as well as audiences at the Universities of Toronto, Macerata, Bonn, Bielefeld, Konstanz, Dubrovnik and Pittsburgh. This research was supported by an Ontario Graduate Research scholarship, the Germany/Europe fund from the University of Toronto, and a postdoctoral fellowship at the Center for Philosophy of Science at the University of Pittsburgh.

References

Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.
Arthur, R. (1999). On thought experiments as a priori science. *International Studies in the Philosophy of Science, 13*, 215-229.
Bangu, S. (2015). Scientific progress, understanding and unification. In I. D. Toader, G. Sandu, & I. Pârvu (Eds.), *Boston studies in the philosophy and history of science: 313. Romanian studies in philosophy of science* (pp. 239-253).
Beller, M. (1999). *Quantum dialogue*. Chicago: University of Chicago Press.
Bird, A. (2007). What is scientific progress? *Noûs, 41*, 64-89.
Bird, A. (2008). Scientific progress as accumulation of knowledge: A reply to Rowbottom. *Studies in History and Philosophy of Science, 39*, 279-281.
Bishop, M. (1999). Why thought experiments are not arguments. *Philosophy of Science, 66*, 534-541.
Bohr, N. (1949). Discussion with Einstein on epistemological problems in atomic physics. In P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-Scientist* (pp. 201-241). Cambridge: Cambridge University Press.
Brown, J. (1993). Why empiricism won't work. *Proceedings of the Philosophy of Science Association, 2*, 271-279.
Brown, J. (2004). Why thought experiments do transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in the philosophy of science* (pp. 23-43). Malden, MA: Blackwell.
Camilleri, K. (2007). Indeterminacy and the limits of classical concepts: The transformation of Heisenberg's thought. *Perspectives on Science, 15*, 178-201.
Camilleri, K. (2014). Towards a constructivist epistemology of thought experiments. *Synthese, 191*, 1697-1716.
Cassidy, D. (1992). *Uncertainty, the life and science of Werner Heisenberg*. New York: Freeman.
Chang, H. (2009). Ontological principles and the intelligibility of epistemic activities. In H. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding* (pp. 64-82). Pittsburgh: University of Pittsburgh Press.
Dancy, J. (1985). The role of imaginary cases in ethics. *Pacific Philosophical Quarterly, 66*, 141-153.
Darwin, C. (1859/1872). *On the origin of species*. Online valorium edition. In B. Bordalejo (Ed.). http://darwin-online.org.uk/Variorum/1859/1859-1-dns.html Accessed 19.10.15.
Dellsén, F. (2016). Scientific progress: Knowledge versus understanding. *Studies in History and Philosophy of Science, 56*, 72-83.
De Regt, H. (2009). The epistemic value of understanding. *Philosophy of Science, 76*, 585-597.
De Regt, H., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese, 144*, 137-170.
De Regt, H., Leonelli, S., & Eigner, K. (Eds.). (2009). *Scientific understanding: Philosophical perspectives*. Pittsburgh: University of Pittsburgh Press.
Duhem, P. (1954). *The aim and structure of physical theory*. Princeton: Princeton University Press.
Elgin, C. (1993). Understanding art and science. *Synthese, 95*, 13-28.
Elgin, C. (1996). *Considered judgment*. Princeton: Princeton University Press.
Elgin, C. (2004). True enough. *Philosophical Issues, 14*, 113-121.
Elgin, C. (2006). From knowledge to understanding. In S. Hetherington (Ed.), *Epistemology futures* (pp. 199-215). Oxford: Oxford University Press.
van Fraassen, B. (1980). *The scientific image*. Oxford: Clarendon.
van Fraassen, B. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press.
Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*, 5-19.
Garber, E., Brush, S., & Everitt, C. W. F. (Eds.). (1996). *Maxwell on heat and statistical mechanics: On avoiding all personal enquires of molecules*. Cranbury, NJ: Associated University Presses.
Gendler, T. (2004). Thought experiments rethought—and reperceived. *Philosophy of Science, 71*, 1152-1163.
Gendler, T. S. (1998). Galileo and the indispensability of scientific thought experiments. *British Journal for the Philosophy of Science, 49*, 397-424.
Gendler, T. S. (2000). *Thought experiment: On the powers and limits of imaginary cases*. New York: Garland Press.
Gooding, D. (1993). What is experimental about thought experiments? *PSA 1992, 2*, 280-290.
Gooding, D. (1994). Imaginary science. *British Journal for the Philosophy of Science, 45*, 1029-1045.
Grimm, S. (2006). Is understanding a species of knowledge? *British Journal for the Philosophy of Science, 57*, 515-535.
Grimm, S. (2012). The value of understanding. *Philosophy Compass, 7*, 103-117.
Häggqvist, S. (1996). *Thought experiments in philosophy*. Stockholm: Almqvist & Wiksell International.
Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge: MIT Press.
Heisenberg, W. (1925). Über quantentheoretische Umdeutung kinematischer und Mechanischer Beziehungen. *Zeitschrift für Physik, 33*, 879-983.
Heisenberg, W. (1930). *The physical principles of the quantum theory* (C. Eckhart and F. C. Hoyt, Trans.). Chicago: University of Chicago Press
Heisenberg, W. (1984). Gesammelte Werke. In W. Blum (Ed.), *Hans-Peter Dürr and Helmut Rechenberg*. München: Piper.
Hilgevoord, J. (1998). The uncertainty principle for energy and time. II. *American Journal of Physics, 66*, 396-402.
Hilgevoord, J. (2006). The uncertainty principle. *Stanford Online Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/qt-uncertainty Accessed 19.10.15.
Hnizdo, V. (2002). On Bohr's response to the clock-in-the-box thought experiment of Einstein. *European Journal of Physics, 23*, L9-L13.
Humphreys, P. (1993). Seven theses on thought experiments. In J. Earman (Ed.), *Philosophical problems of the internal and external world: Essays on the philosophy of Adolf Grünbaum* (pp. 205-227). Pittsburgh: University of Pittsburgh Press.
Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*, 507-531.
Kitcher, P. (2002). Scientific knowledge. In P. K. Moser (Ed.), *The Oxford Handbook of Epistemology* (pp. 385-407). New York: Oxford University Press.
Kosso, P. (2006). Scientific understanding. *Foundations of Science, 12*, 173-188.
Kragh, H. (2002). *Quantum generations: A history of physics in the twentieth century*. Princeton: Princeton University Press.
Kudaka, S., & Matsumoto, S. (1999). Uncertainty principle for proper time and mass. *Journal of Mathematical Physics, 40*, 1237-1245.
Kuhn, T. S. (1977). A function for thought experiments. In *The essential tension* (pp. 240-265). Chicago: University of Chicago Press.
Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. New York: Cambridge University Press.
Kvanvig, J. (2009). The value of understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 95-111). New York: Oxford University Press.
Lipton, P. (2004). *Inference to the best explanation*. New York: Routledge.
Lipton, P. (2009). Understanding without explanation. In H. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding* (pp. 43-63). Pittsburgh: University of Pittsburgh Press.
Machamer, P., & Woody, A. (1994). A model of intelligibility in science: Using Galileo's balance as a model for understanding the motion of bodies. *Science and Education, 3*, 215-244.
Maxwell, J. C. (1860a). On the dynamical theory of gases. Reprinted in *British Association* 9. Read: Aberdeen, September 1859. NLS: J. 198-199.
Maxwell, J. C. (1860b). 'Illustrations of the dynamical theory of gases.' I. On the motion and collision of perfectly elastic spheres. *Philosophical Magazine, XIX*, 19-32.

Maxwell, J. C. (1860c). 'Illustrations of the dynamical theory of gases.' II. On the process of diffusion of two or more kinds of moving particles among one another. *Philosophical Magazine, XX*, 21-33.

Maxwell, J. C. (1860d). 'Illustrations of the dynamical theory of gases.' III. On the collision of perfectly elastic bodies of any form. *Philosophical Magazine, XX*, 33-37.

Maxwell, J. C. (1861). On the results of Bernoulli's theory of gases as applied to their internal friction, their diffusion, and their conductivity for heat. *Rep. Brit. Assoc.*, 15-16. Read, Oxford, June & July. NLS: J. 198-199.

Maxwell, J. C. (1866). On the viscosity or internal friction of air and other gases. Bakerian Lecture. *Philosophical Transactions of the Royal Society of London, XV CLVI*, 249-268.

Maxwell, J. C. (1871). *A theory of heat*. London: Textbooks of Science.

Meinong, A. (1907)1973. Das Gedankenexperiment. In R. Haller and R. Kindinger (Eds.), *Über die Stellung der Gegenstandstheorie im System der Wissenschaften* (pp. 273-283). Graz-Austria: Akademische Druck und Verlagsanstalt.

Mott, N., & Peierls, R. (1977). Werner Heisenberg. *Biographical Memoirs of Fellows of the Royal Society, 23*, 213-251.

Myrvold, W. (2011). Statistical mechanics and thermodynamics: A Maxwellian view. *Studies in History and Philosophy of Science Part B, 42*, 237-243.

Nersessian, N. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Cognitive models of science* (pp. 3-44). Minneapolis: University of Minnesota Press.

Nersessian, N. (2007). Thought experiments as mental modelling: Empiricism without logic. *Croatian Journal of Philosophy, VII*, 125-161.

Nobel Prize in Physics, 1932. http://www.nobelprize.org/nobel_prizes/physics/laureates/1932/index.html Accessed 19.10.15.

Norton, J. (2004). Why thought experiments do not transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in the philosophy of science* (pp. 44-66). Somerset: Wiley-Blackwell.

Norton, J. (Unpublished). *Einstein for everyone*. University of Pittsburgh. http://www.pitt.edu/~jdnorton/teaching/HPS_0410/chapters_2013_Jan_1/index.html Accessed 19.10.15.

Norton, J., & Earman, J. (1999a). Exorcist XIV: The wrath of Maxwell's demon. Part I. From Maxwell to Szilard. *Studies in the History and Philosophy of Physics, 29*, 435-471.

Norton, J., & Earman, J. (1999b). Exorcist XIV: The wrath of Maxwell's demon. Part II. From Szilard to Landauer and beyond. *Studies in the History and Philosophy of Physics, 29*, 435-471.

Pais, A. (1982). *Subtle is the lord…*. Oxford: Clarendon Press.

Pritchard, D. (2009). Knowledge, understanding, and epistemic value. In A. O'Hear (Ed.), *Epistemology* (pp. 19-43). New York: Cambridge University Press.

Pritchard, D. (2010). Knowledge and understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *The nature and value of knowledge: Three investigations* (pp. 3-90). New York: Oxford University Press.

Saint-Exupéry, A. de (1943/2013). *The Little Prince*. New York: Mariner Books.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.

Stuart, M. (2014). Cognitive science and thought experiments: A refutation of Paul Thagard. *Perspectives on Science, 22*, 264-287.

Stuart, M. (2016). Imagination: A sine qua non of science. *Croatian Journal of Philosophy* (forthcoming)

Thagard, P. (2010). *The brain and the meaning of life*. Princeton: Princeton University Press.

Thagard, P. (2014). Thought experiments considered harmful. *Perspectives on Science, 22*, 288-305.

de la Torre, A. C., Daleo, A., & Garcia-Mata, I. (2000). The photon-box Bohr-Einstein debate demythologized. *European Journal of Physics, 21*, 253-260.

Toulmin, S. (1972). *Human understanding*. Princeton: Princeton University Press.

Treder, H.-J. (1970). The Einstein-Bohr box experiment. In W. Yourgrau, & A. van der Merwe (Eds.), *Perspectives in quantum theory*. Cambridge: MIT Press.

Treder, H.-J. (1975). The Einstein shift in Einstein's box experiment. *Foundations of Physics, 5*, 135-142.

Van Dyck, M. (2003). The roles of one thought experiment in interpreting quantum mechanics: Werner Heisenberg meets Thomas Kuhn. *Philosophica, 72*, 79-103.

Wheeler, J., & Zurek, W. H. (Eds.). (1983). *Quantum theory and measurement*. Princeton, NJ: Princeton University Press.

Wilkes, K. (1988). *Real people: Personal identity without thought experiments*. Oxford: Oxford University Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Zagzebski, L. (2001). Recovering understanding. In M. Steup (Ed.), *Knowledge, truth, and duty: Essays on epistemic justification, responsibility, and virtue* (pp. 234-251). New York: Oxford University Press.