Thought Experiments in Science

by

Michael T. Stuart

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Institute for the History and Philosophy of Science and Technology
University of Toronto

© Copyright by Michael T. Stuart 2015
Thought Experiments in Science

Michael T. Stuart

Doctor of Philosophy

Institute for the History and Philosophy of Science and Technology
University of Toronto

2015

Abstract

Thought experiments are a means of imaginative reasoning with an employment record longer than two and a half thousand years. Used by Aristotle, Galileo, Newton, Darwin, Maxwell, and Einstein, they form part of the education of every scientist alive today.

While most scientific instruments aim to increase the precision of our interaction with empirical data, thought experiments leave the empirical realm behind. They spin buckets in an empty universe, summon demons to play with particles, and challenge us to throw spears at the edge of space. If tools of the imagination like thought experiments are important in science, what role are they playing?

Using the methods of history, philosophy and cognitive science, I argue that while thought experiments do not always discover or justify new facts, they do usually increase the empirical content of theoretical structures (laws, models, concepts, etc.) for an agent. Empirical content is increased when the agent connects a theoretical structure to existing concepts, experiences, values and abilities. Through these connections, thought experiments increase scientific understanding. Knowledge can be produced when understanding is applied, but this is a separate achievement. I argue that the understanding produced by thought experiments is fallible but
necessary for scientific progress, in the same way that intuitions about what a speaker means are fallible but necessary for linguistic interpretation.

Chapter 1 introduces the philosophical literature on thought experiments. Chapters 2-4 reject two important accounts of thought experiments. Chapter 5 examines historical cases to focus on the role of thought experiments in science. Chapter 6 considers results in cognitive and social science. Chapter 7 presents a transcendental argument that grounds our ability to increase empirical content through thought experiments in human imagination, and explores some implications of this argument for the relationships among imagination, understanding, explanation, knowledge, representation and objectivity.
Acknowledgments

Yiftach Fehige and Jim Brown have attempted to forge a philosopher out of a graduate student, and I have been lucky to have them as hammer and anvil. I thank Yiftach foremost for his time, advice, enthusiasm, friendship, and for caring enough to be harsh. I am grateful to Jim Brown for his unceasing dedication, asking only careful reflection and clever argument (I hope I have provided a little of each). I also thank Jossi Berkovitz, who has been a very helpful and friendly critic.

There are many other professional philosophers to whom I owe gratitude. Marco Buzzi is a fountain of knowledge who I will always try to emulate. He is humble, open-minded, generous, and carefully balanced between analytic and continental. I am also very grateful to Catherine Elgin for doing me the honour of reading and commenting on my thesis. I admire her greatly, and was very lucky to have had her as an external committee member. Thank you also to Brian Baigrie, Elke Brendel, Martin Carrier, Sören Häggqvist, Stephan Hartmann, Hans Radder, Tobias Rosefeldt, Bill Seager, Paul Thagard, Marcel Weber, and Kenneth Westphal, whose comments and conversations have given me much to think about. Denis Walsh has been endlessly patient with me, and always inspiring. Thank you also to Chen-Pang Yeang, Craig Fraser, Margie Morrison and Anjan Chakravartty. Denise Horsley and Muna Salloum have been incredibly supportive during my time at the University of Toronto, making everything easier and warmer. At York University, Jagdish Hattiangadi lit a fire that burns still, and Judy Pelham has been a lasting influence and friend.

Equally important is the support I have received from my colleagues: Curtis Forbes, Paul Greenham, Cory Lewis, Greg Lusk, Geordie McComb, Eran Tal, Harald Wiltsche, Aaron Wright, and Jai Virdi, many of whom have read this thesis almost as many times as I have. I hope to work with each of you again. Thank you all.
Table of Contents

List of Figures ........................................................................................................... vii

Chapter 1 The Paradox of Thought Experiments ......................................................... 1
1  The Study of Thought Experiments ........................................................................ 2
2  Paradigms and Paradox ....................................................................................... 4
3  Analyzing the Paradox ....................................................................................... 9
   3.1  Novelty ........................................................................................................ 9
   3.2  Empirical Import .......................................................................................... 10
      3.2.1  Empirical Import = Empirical (Semantic) Content ......................... 12
      3.2.2  Empirical Import = Empirical Information ...................................... 14
      3.2.3  Empirical Import = Empirical Evidence ......................................... 17
   3.3  Inside the Head ............................................................................................ 19
4  Types of Account: Cartesian and Baconian ......................................................... 22
5  Above Distinctions Applied to the Modern Epistemological Accounts of Thought Experiments ........................................................................................................... 23
6  Chapter Summaries ............................................................................................ 31

Chapter 2 Thought Experiments and Arguments ..................................................... 36
1  Some Context ..................................................................................................... 37
2  The Kinds of Inferences that Count as Arguments ............................................. 39
3  Identifying Arguments ....................................................................................... 42
4  Performing and Justifying Thought Experiments ............................................. 44
5  The Move to Logic ............................................................................................. 46
6  A “Very Generalized Logic” ............................................................................ 47
7  A Broader Conception of Argument .................................................................. 48
8  Conclusion ......................................................................................................... 53

Chapter 3 Thought Experiments and Logic .............................................................. 54
1  The Mapping Account of the Applicability of Mathematics ............................... 57
2  The Mapping Account Applied to Logic ............................................................. 59
3  A Case Study: Operational Constraints ............................................................. 64
4  Implications for the Argument View .................................................................. 66
5  Conclusion ......................................................................................................... 71

Chapter 4 Thought Experiments and Cognitive Science ........................................ 73
1  Thagard’s Phenomenology of Thought Experiments ....................................... 74
2  Against Thagard’s Phenomenology of Thought Experiments ......................... 77
   2.1  Thought Experiments as Independent of Experience .............................. 78
   2.2  Thought Experiments as a Means to Necessary Truth ......................... 79
   2.3  Against Thagard’s Definition of Thought Experiments ......................... 82
   2.4  The Method of Thought Experiments .................................................... 83
   2.5  From Conceivability to Possibility ......................................................... 84
   2.6  Thought Experiments as Weakly Evidential ......................................... 86
   2.7  Thought Experiments as Strongly Evidential ....................................... 86
   2.8  Evidence ..................................................................................................... 88
3  Cognitive Science, Thought Experiments and Mental Models ......................... 89
4  Conclusion ......................................................................................................... 93

Chapter 5 Five Thought Experiments Reconsidered ............................................... 96
1  Maxwell’s Demon ............................................................................................. 97
2  The Clock in the Box ....................................................................................... 108
3  Einstein’s Elevator ......................................................................................... 123
4  Heisenberg’s Microscope ............................................................................... 128
5  Darwin’s Vertebrate Eye ............................................................................... 132
6  Understanding Theoretical Structures ............................................................ 137
List of Figures

Fig. 1: How a Thought Experiment Might Produce Something Novel ........................................... 9
Fig. 2: Some Senses of Empirical Import ...................................................................................... 11
Fig. 3: How a Thought Experiment Might be “Inside the Head” .................................................. 19
Fig. 4: (Some of) the Conceptual Space of the Paradox ................................................................. 21
Fig. 5: Euler Circles to Reason through a Claim ............................................................................ 65
Fig. 6: Convolution Occurring In Simulated Neurons ................................................................. 91
Fig. 7: Einstein’s Clock in the Box, Depicted by Bohr ................................................................. 110
Fig. 8: Bohr’s Pseudo-Realistic Drawing of the Clock in the Box .............................................. 110
Chapter 1
The Paradox of Thought Experiments

Thought experiments (at least in some cases) allow us to intuit laws of nature. Intuitions, remember, are nonsensory perceptions of abstract entities. Because they do not involve the senses, they transcend experience, and give us a priori knowledge of the laws of nature.

(Brown 2004, 34)

If this can be taken at face value, thought experiments perform epistemic magic.

(Norton 2004b, 44)

Thought experiments can be found in almost all disciplines of human inquiry, going back at least two and a half millennia (Rescher 1991). Partially responsible for this ubiquity is the flexibility of the human imagination. We have no difficulty imagining features of mathematical, political, moral, biological, physical or metaphysical systems. And this should be expected, as the human mind is at least partially responsible for the concepts that define those disciplines. In the last 40 years a great deal of work has been done on the power of the imagination, whether characterized as the ability to conceive (Gendler and Hawthorne 2002), reason counterfactually (Byrne 2005, Kahneman and Miller 1986, Lewis 1973, Mandel et al. 2014), simulate mentally (Khemlani et al. 2013, Markman et al. 2009), or produce mental imagery (Kosslyn 1994, Kosslyn et al. 2006, Pylyshyn 2002). But little has focused specifically on the epistemological role of the imagination in scientific thought experiments. In this thesis I will argue that performing a thought experiment can trigger the imagination in a way that increases the empirical content of a theoretical structure (proposition, model, concept, etc.) for an agent. This has broader implications for the role of the imagination in science, generally.

I begin with a brief discussion of thought experiments themselves, before turning to the literature that studies them.
1 The Study of Thought Experiments

While there is no standard definition for thought experiments, “we recognize them when we see them” (Brown 1991b, 122). A short enumeration of some classic thought experiments displays just how interesting and diverse they can be. Examples include Maxwell’s demon, Einstein’s elevator (and train, and stationary lightwave), Schrödinger’s cat, Searle’s Chinese room, Putnam’s twin Earth (and brains in vats), Nozick’s experience machine, Rawl’s original position, Newton’s bucket (and cannonball), Heisenberg’s microscope, Jackson’s colour scientist, Thomson’s violinist, Chalmers’s zombies, Wittgenstein’s beetle, the Prisoner’s Dilemma, Plato’s cave (and ring of Gyges), Lucretius’s throwing a spear at the edge of the universe, Quine’s gavagai, Davidson’s Swampman, Eddington’s monkeys who type Hamlet, Stevin’s chain draped over a prism, Poincaré’s diskworld, and Foot’s trolley problem. All of these can be found in collections of thought experiments (Tittle 2004 and Cohen 2004), and there is now a textbook that introduces students to philosophy entirely through thought experiments (Schick and Vaughn 2012).

Several of the above-mentioned thought experiments have taken on lives of their own, despite Hacking’s (1992) claim that they cannot do this. One example is Philippa Foot’s “trolley problem” (1967, which has been revised, reappropriated and altered by (in chronological order): Thomson (1976, 1985), Unger (1996), Kamm (1989), Singer (2005), Navarrete et al. (2012), and Cathcart (2013). Another example is Maxwell’s demon (Maxwell 1870), which has been criticized, endorsed and enhanced by (in chronological order): Brillouin (1951), Daub (1970), Heimann (1970), Zurek (1984), Collier (1990), Maddox (1990), Zhang and Zhang (1992), Earman and Norton (1998) and Leff and Rex (2002).

Many famous works of art have been characterized as thought experiments, including Huckleberry Finn, To Kill a Mockingbird, Oedipus Rex, A Tale of Two Cities, Lolita, Middlemarch, The Matrix, 2001: A Space Odyssey, Henry V, King Lear, Hamlet, Animal Farm, and Uncle Tom’s Cabin (Elgin 2014).
Thought experiments bespeckle scientific texts like Galileo’s *Discourse*, Newton’s *Principia*, and Darwin’s *Origin*, and they are equally common in pure and applied mathematics, where they are central in research from geometry (Lakatos 1976) to infinity (Galilei 1638, 32; Hilbert 2013).

Of course, a thought experiment need not be historically important; we can invent them at will. Tamar Gendler describes a thought experiment in which we imagine our next-door neighbour’s living room with an elephant in it, and then ask if there would be enough room left to ride a bicycle without tipping over (Gendler 2004, 1156-1157).

Thought experiments clearly form an extremely diverse set of mental activities. How should they be investigated? First, we have to recognize that their being easy to perform does not mean they will be easy to understand. Expecting to understand thought experiments due to our long experience using them is like expecting birds to understand aerodynamics because they can fly. Yet while birds do not need a theory of aerodynamics, we do need a philosophical account of thought experiments. This is because thought experiments sometimes go wrong, and it is not always obvious why. Laws of aerodynamics are stable and the bodies and instincts of birds have evolved to take advantage of them. Our instincts concerning what we should infer from an imaginary scenario are not so reliable. Using our imagination to figure out the behaviour of subatomic particles might be like a bird trying to fly in outer space.

To proceed, we need more than mere introspection. One important source of information is history. Obviously, we could not deduce a philosophical account of thought experiments from a historical enumeration of “successful” thought experiments because such an enumeration would already presuppose many philosophical assumptions concerning what thought experiments are, what they can do, and when they should be counted successful. Still, history is absolutely crucial. For one thing, historical work on thought experiments such as Gellard (2011), Ierodiakonou (2005, 2011), Knuuttila and Kukkonen (2011), Kühne (2005), Lautner (2011), and Palmerino (2011), has already prompted many important philosophical issues that might have otherwise gone unnoticed. For instance, there are features of thought experiments that are common to some periods and not others. Ancient Greeks were comfortable employing impossible premises in their thought experiments (Ierodiakonou 2011), while many modern writers are not (Wilkes 1988 1-48, and see Brown and Stuart 2013).
Comparative historical study prompts other philosophical questions, such as: are there features common to all thought experiments? Do different communities draw different lines between thought experiments, fictions, models and arguments? How contextual are the success criteria for thought experiments, and what causes a community to change them? Answering these questions requires careful historical work.

Aside from historical and philosophical methods there are also sociological methods (including ethnography) and psychological methods (including the wide range of methods employed in cognitive science). Each of these provides more information from which a fully-informed account of thought experiments must draw.

This dissertation attempts to benefit from each of these sources. Chapters 2-4 will employ philosophical argument. Chapter 5 relies on historical case studies, and Chapter 6 turns to social and cognitive science. This order purposely mirrors the order of investigation that has played out over the last 30 years, which I will now present. Instead of going all the way back to the Presocratics (Rescher 1991), Plato (Miščević 2012), Descartes and Hume (Gendler and Hawthorne 2002), the German idealists (Buzzoni forthcoming, Kühne 2005, Fehige and Stuart 2014) or Mach (Sorensen 1992), I begin with Thomas Kuhn.

2 Paradigms and Paradox

In my opinion, Kuhn set the focus for the current period of discussion concerning thought experiments. Unlike Mach and those before him, Kuhn wrote almost exclusively about thought experiments as a tool for motivating or justifying claims during scientific revolutions. For Kuhn, “A crisis induced by the failure of expectation and followed by revolution is at the heart of the thought-experimental situations we have been examining. Conversely, thought experiment is one of the essential analytic tools which are deployed during crisis and which then help to promote basic conceptual reform” (Kuhn 1977, 263). He cites Einstein’s train, Heisenberg’s microscope, and several fragments from Galileo as examples of thought experiments that play this role in theory change. He calls these “an important class of thought experiments” (260-261), and he concludes that “from thought experiments most people learn about their concepts and the world together” (253).
For Kuhn, revolutionary thought experiments are not used to generate new facts, but to ease us through the irrational period of crisis that exists between scientific paradigms, guiding us back to the rational progress of what Kuhn calls “normal” science. In a period of crisis, we must weigh the competing claims, methods and promises of rival paradigms, and it seems that thought experiments help us (partially) transcend the confines of paradigms, which is necessary if we are to be convinced of a new world-view. It is then argued that by changing world-views, we learn about the world.

Kuhn’s answer to the question of how thought experiments fuel scientific progress did not win widespread acceptance, although there is some sympathy (for example, Sorensen 1992, Gendler 1998 and Van Dyck 2003). What I would like to draw attention to is the importance of Kuhn’s idea that thought experiments play a justificatory role in science, and especially in scientific revolutions. This idea was recognized by those who organized the first conference on thought experiments in 1986, and it has been a focal point of the discussion ever since. The proceedings of the conference were published in Horowitz and Massey (1991), and on the first page of the introduction the editors point out that what is at stake is a paradox inspired by Kuhn’s paper, which they called the “paradox of thought experiments.” It consists in the “puzzling fact that thought experiments often have novel empirical import even though they are conducted entirely inside one’s head.”

This wording is pretty close to the way Kuhn framed the problem, although not exactly. In Kuhn’s words the problem is: “How, then, relying exclusively upon familiar data, can a thought experiment lead to new knowledge or to a new understanding of nature?” (1977, 241).

Is this really a paradox? W.V.O. Quine defines a paradox as “any conclusion that at first seems absurd, but that has an argument to sustain it” (1966, 3). Piotr Łukowski provides a similar definition: a paradox is “a thought construction, which leads to an unexpected contradiction” (2011, 1). Doris Olin agrees: “a paradox is an argument in which there appears to be correct reasoning from true premises to a false conclusion” (2003, 6). According to these conceptions of paradox, the puzzle about thought experiments is probably not a paradox. However, according to more inclusive conceptions, it is. Roy Sorensen defines paradoxes as “conflicting, well-credentialed answers” to problems (2011), and elsewhere as questions “that suspend us between
too many good answers” (2003, xii). Sorensen regards paradoxes “as the atoms of philosophy because they constitute the basic points of departure for disciplined speculation” (2003, xi). According to Sorensen’s characterization at least, the puzzle about thought experiments does indeed become a paradox, but only in the years that follow Kuhn’s paper.

The transition from puzzle to paradox takes place when Kuhn’s open-ended question transforms into a dilemma between two options: a world with epistemic magic, and one without. To see how this happens we have to look at the work of James R. Brown and John D. Norton. Each writer assumes with Kuhn that thought experiments can play a justificatory role in scientific revolutions. Brown presents a sketch of a Platonic theory of thought experiments (1986) and Norton develops a view that is Brown’s foil; an empirical account that characterizes thought experiments as arguments (1991). Each of these are attempts to resolve Kuhn’s puzzle, although only Brown makes the connection to Kuhn explicit.

Brown begins with a thought experiment from Galileo in which “we have a transition from one theory to another which is quite remarkable. There has been no new empirical evidence. The old theory was rationally believed before the thought experiment, but was shown to be absurd by it. The thought experiment established rational belief in a new theory” (1986, 10). It does this a priori, which for Brown, means independent of experience. Brown argues that it is a priori for five reasons, of which I will mention three. One is that “there has been no new observational data” (11). Another is that “it is not a case of seeing old empirical data in a new way” (11). (Brown writes, “This is essentially Kuhn’s thesis” 1986, 11). And the third reason is logical:

Galileo has not merely deduced his theory of free-fall from already given empirical premisses. Nor is his achievement to be trivialized by saying it follows from the contradiction in Aristotle’s account. If that were all that is going on then Galileo could also have deduced ‘The moon is made of green cheese,’ all of the quantum theory, and anything else he liked. Moreover, Galileo’s theory is not a formal truth that one could have inferred from no premisses at all because it says nothing about the world. That is, it is not some sort of analytic truth. Rather, it is synthetic a priori. (Brown 1986, 11-12)

While Brown rejects Kuhn’s exclusionary focus on revolutionary thought experiments, he accepts that at least some perform the role that Kuhn envisaged, “a crucial role in paradigm
change” (Brown 1986, 2). They play this role by providing reasons to reject one theory and adopt another, and those reasons are not strictly logical, nor do they rely solely on previous sense-experience.

Norton appears to agree with Brown concerning the problem:

Thought experiments in physics provide or purport to provide us information about the physical world. Since they are *thought* experiments rather than *physical* experiments, this information does not come from the reporting of new empirical data.” But he draws a very different conclusion from this state of affairs: “Thus there is only one non-controversial source from which this information can come: it is elicited from information we already have by an identifiable argument...The alternative to this view is to suppose that thought experiments provide some new and even mysterious route to knowledge of the physical world. (1991, 129)

Norton hereby presents Kuhn’s puzzle in the form of a dichotomy, which is what makes it paradoxical. It becomes a question with conflicting but well-credentialed answers. Given that thought experiments provide or purport to provide information about the physical world, yet do not require new information about the physical world, either the new information is a rearrangement of old data, or else it comes from rational insight.

thought experiments mention the paradox in the abstract or introduction. Interestingly, if this were 2009, the same comparison would tell us that sixty-nine percent of the relevant literature mentioned the paradox in the abstract or introduction. This difference is probably due to the influx of papers written since 2009 on the descriptive psychology of thought experiments. The percentage would be even higher if the search was extended beyond the abstract and introduction, and to monographs as well. (We could then include Brown 1991a; Buzzoni 2008; De Mey 2005, 2006b; Gendler 2000; Georgiou 2007; Häggqvist 1996; Sorensen 1992; and others).

What is important is that most of these above-listed contributions present Kuhn’s problem in slightly different terms, or call it by a different name. For example Norton refers to it as “the epistemological problem of thought experiments in the sciences” (2004a, 1139), while it is called the “problem of informativeness” for “scientific thought experiments of evidential significance” by Brendel (2004, 89) and Fehige (2013, 56). It is still called the “Fundamental Paradox of Thought Experiments” by Clement (2002, 32; 2003, 261). Nevertheless, as the accounts are framed in contrast to one another and the definite article is almost always used, I take it that the parties to the debate assume it is the same problem they are addressing.

Given that the paradox is rarely presented in the same words, we should ask whether there is really only one main paradox. In the next section I map out the conceptual space of the paradox and show that it admits of some very different interpretations. Outlining these interpretations provides an easy way to sort out the rapidly expanding literature on thought experiments (which I do in section 5, below), and it also shows that there are viable ways to formulate the paradox that no one pursues. After my case studies in Chapter 5, I will identify a role for thought experiments in science that is explainable only if we adopt one such characterization of the paradox. A consequence is that the account developed will not necessarily conflict with other accounts in the literature, because it asks a different question.

I begin my exploration of the paradox with the formulation of Horowitz and Massey because these authors characterize it in the context of introducing the results of the first conference on thought experiments. Because of this, they chose a statement of the paradox they hoped would cover what was interesting in all the contributions.
3 Analyzing the Paradox

Horowitz and Massey characterize the paradox as the “puzzling fact that thought experiments often have novel empirical import even though they are conducted entirely inside one’s head.” There are three features of this statement that deserve pause: “novel,” “empirical import,” and “entirely inside one’s head.” As we will see, all three of these features are present in each of the formulations in the literature, and they are often understood differently.

3.1 Novelty

There are several ways the outcome of a thought experiment can be considered novel. Here are nine:

![Fig. 1: How a Thought Experiment Might Produce Something Novel](image)

---

1 For this and the following three figures the size of the segments are not intended to be relevant. I only want to display the options.
One way a thought experiment can be considered novel is when its outcome is surprising. I will call this “psychological” novelty.

Another sense of novelty concerns the derivability of the outcome of a thought experiment. There are different strengths of this sense of novelty, for example, depending on whether it is the average person who cannot derive the conclusion, or some ideal logical agent. The strength of this sort of novelty also depends on the sources from which the supposed derivation should be made. Possible sources include stated premises, background knowledge, sense-experience, modal intuition, and many other things. When different sources are considered, different outcomes of thought experiments will count as novel. For example, if we limit ourselves to the explicit premises and background theory of Galileo’s falling body thought experiment, then Brown is probably correct that its outcome is not derivable, and is therefore novel. However if we present an ideal reasoner with all possible sources, perhaps its outcome is derivable.

More senses of novelty emerge from considering the output of thought experiments. Perhaps a thought experiment causes us to acquire a new belief. Before the experiment we did not assent to a given proposition, but now we do. Or perhaps it provides us with a new experience, in the sense that it exposes us to a representation of an event or phenomenon that we were not exposed to before performing the thought experiment. Or perhaps what is novel is a change in the epistemological status of a belief. In this sense we gain or lose knowledge (or understanding) as a result of the thought experiment. Or perhaps it establishes a new relationship among existing beliefs, which might be expressed in the form of new logical or psychological ties between propositions or concepts. Or perhaps we emerge with a new valuation or emotional connection. Or perhaps a thought experiment gives us a new ability, in the sense that previously unachievable goals become achievable after the thought experiment has been performed.

Philosophers have invoked most of these senses of novelty, and as we will see below, most invoke more than one.

3.2 Empirical Import

There are many ways to expand the notion of “empirical import,” some of which appear in Figure 2. Not all of these senses of empirical import can be provided by a thought experiment. In
the most general sense, empirical import only implies *relevance* to sense-experience. To increase specificity we must interpret this relevancy requirement. I will consider three options: to be relevant to experience means 1) to increase or change the empirical (semantic) *content* of something, 2) to increase or change our empirical *information* about something, or 3) to increase or change our empirical *evidence* for something. As with the different senses of novelty, philosophers combine these senses of empirical import.

![Fig. 2: Some Senses of Empirical Import](image-url)
3.2.1 Empirical Import = Empirical (Semantic) Content

Empirical import understood as empirical content could be interpreted in at least the following three ways. It could be a) the pairing or mapping of terms to their extra-mental referents, b) the empirical psychological “sense” of something, or c) a list of what would need to be the case in the extra-mental world for the conclusion of a thought experiment to be true.

According to a), thought experiments would help us map representations to phenomena. For example, Edmund Gettier presented a famous thought experiment against the view that knowledge is justified true belief (Gettier 1963). He did this by crafting a possible scenario in which someone has a true justified belief that is not knowledge. If the thought experiment is successful, it shows that we should not map instances of the concept KNOWLEDGE necessarily to instances of true justified belief. How thought experiments establish or alter mappings from concepts or propositions to the world (with rational ostentation or at least without empirical ostentation) is one way to phrase the paradox about thought experiments.

According to b), thought experiments affect the empirical psychological content of a concept, term, model, etc. What it means to have empirical psychological content is an issue that is related to foundational theories of meaning. Empirical psychological content might be meaning, connotation, intension, a propositional attitude, a non-propositional mental representation, or whatever connects a representation and its extension. I will not be arguing for one of these views. For the moment, I merely want to ask how psychological content could be empirical. One way is for it to enjoy a causal tie to the extra-mental world. Perhaps my psychological content for DOG is caused by various experiences with wagging tails and wet-noses. Causal interaction is therefore one way psychological content could be empirical.

But the empirical element of psychological content could come from something else, like having at least one element of the content refer to something we have personally experienced. In this sense, ∞ would have less empirical content than BASEBALL, since I have seen baseball played but

\footnote{In this thesis I use SMALL CAPS to mark concepts.}
I have never seen anything infinite. Alternatively, we could say that psychological content is empirical when at least one of its elements refers to something that is experience-able.

Let me apply this to a thought experiment. Consider Einstein’s train. In this thought experiment, we begin with the experience of a stationary observer witnessing a train moving close to the speed of light. For two flashes of light—one on each end of the train—to appear simultaneously to the stationary observer when the train passes directly in front of her, the flashes must be ignited at different times (the flash at the front of the train must be ignited later than the flash at the rear). For the flashes to appear simultaneously to a passenger on the same train who is sitting in the middle of the train, however, the flashes must go off simultaneously, since the light has to travel equal distances to the center of the train. The conclusion is that events counted as simultaneous will depend on the observer’s frame of reference. We could interpret this thought experiment as changing the empirical content of the concept SIMULTANEITY from absolute to relative. Before Einstein, simultaneity had a certain connotation that was tacitly or explicitly tied to the concept ABSOLUTE, because whatever was simultaneous in one reference frame was simultaneous in all. ABSOLUTE was part of the content; it was how it was understood. After the thought experiment, we adjust the content of the concept so that it reflects RELATIVITY TO REFERENCE FRAME.

Now, this change in content could be empirical because it was inspired by considering representations of phenomena that we have experienced, like trains and flashes of light. Or it could be empirical because SIMULTANEOUS refers to a relationship between extra-mental objects and events. In this sense, the content is empirical because it refers to events we have personally experienced (like synchronizing clocks), or because it refers to events we would expect to experience, although we may never have such experiences (such as recording different chronologies from Earth-based observers).

These different accounts of how psychological content can be empirical do not conflict; we may employ all three at once if we wish. The main point is merely that characterizing empirical import as a change in empirical psychological content is a viable and interesting way to present the paradox as a puzzle concerning how thought experiments affect the sense of a scientific representation.
Finally, a thought experiment could be understood as affecting empirical content by telling us what would need to be true of the world assuming that we are correct concerning the assumptions that figure into the thought experiment. Turning back to the above example, Einstein’s train has empirical content because if it is correct, we should expect to find that travellers moving at very high speeds should experience different orders of events than stationary observers measuring the same events. In this case, the paradox about thought experiments asks how thought experiments can tell us what must be true about the world given our assumptions and an imaginary scenario. This is different from the previous characterization of empirical content because while both involve our expectations about possible experiences, the first portrays thought experiments as affecting the psychological content associated with our representations, where the second portrays thought experiments only as affecting our expectations (and is therefore compatible with an account that eschews meanings or intensions).

Each of these interpretations of the way thought experiments affect empirical content are different. Nevertheless, for each of them, the question of how thought experiments provide novel empirical import will be related to issues in the philosophy of language concerning meaning or reference. One way to address this version of the paradox is therefore to defend a position in the philosophy of language, and use it to explain the cognitive efficacy of thought experiments. Alternatively, we could flip the issue on its head and use research into thought experiments to provide clarity about (or help us to decide between) positions in the philosophy of language. Although valid, these characterizations of the paradox have not been considered much in the literature.

3.2.2 Empirical Import = Empirical Information

Another way to understand empirical import is as empirical information. As before, philosophical divisions arise immediately. I will consider three popular conceptions of information (see Floridi 2010): mathematical, semantic and economical.

Mathematical information is elsewhere referred to as “Shannon information” (introduced in Shannon 1948), and it captures how many “bits” of information are contained in a message. A message may be polluted with noise, making its contents “uncertain.” The measure of the uncertainty of a message’s information is a measure of its informational entropy. A message
which is completely predictable will tell us little that we did not already know, and is defined as having low entropy. A message that is unpredictable will provide more information about the world, and will have higher entropy. A commonly used example is flipping a coin. The 10th flip of a weighted coin will not tell us anything new, because we were already certain it would be heads. The 10th flip of a fair coin, however, will be informative because we could not have predicted this outcome with as much certainty as with the weighted coin. The more informational entropy a message has, the more it tells us about the world, and the less predictable it is from what we already know. This leads to the “inverse relationship principle” (Barwise and Seligman 1997), which states that the informativeness of some information increases as its probability decreases. If the information contained in a message is very probable given current knowledge, it will not tell us much about the world, and will therefore have low entropy. Taken to its conclusion, this creates the “scandal of deduction” (Hintikka 1970): any tautological deduction will have a probability of 1, and will therefore be maximally uninformative. Likewise, since the derivation of any contradiction will have a probability of 0, it will be maximally informative (the second half of the scandal is called the Bar-Hillel Carnap paradox, see Bar-Hillel and Carnap 1952). This is unintuitive, because we know that deductive inferences can be informative, and we generally think that contradictions are uninformative (Heraclitus excepted).

The second (semantic) sense of information is an extension of the mathematical sense (Floridi 2011b). One way to extend information to semantics is simply to identify the amount of meaning in natural language expressions as the amount of (Shannon) information in those expressions. For example, “I am here now” has less informational entropy, is more predictable given what I know, and tells me less about the world, than “it will rain tomorrow.” Adopting this conception of semantic information again appears to provide counterintuitive results when taken to its extreme: given the low probability that a wildly hypothetical statement will be true, hypothetical considerations are more informative than verifiable indicative statements (Dretske 1981, 42). “Frank Sinatra was a hamburger” is more informative than “the beer is in the fridge.” But if we note that such a characterization does not necessarily recommend that thought experiments maximize information, this might be a good way to begin an investigation into how departures from reality usefully increase information.
Floridi wants to avoid a conception of information according to which meaningless jumbles of words are maximally informative, however, so he expands his definition of semantic information to “well-formed, meaningful, and truthful data” (2011a, 31). By adding in a truthfulness condition, he ensures that we do not run into situations where gibberish is more meaningful (carries more information) than sensible expressions. According to this understanding of information, the paradox of thought experiments is a question concerning how novel empirical information, that is, how novel, well-formed, meaningful, and truthful empirical data can be produced from entirely within the head. This is an interesting characterization because hypothetical thought experiments often employ data that are not well-formed (by stretching concepts and breaking grammatical rules) and include false assumptions (like frictionless planes) to generate what we hope are well-formed, meaningful, and truthful data. Trying to show how thought experiments do this could be a very interesting way to address the paradox.

The economic conception of information concerns the value of information to humans living in communities. Certainly thought experiments can change the value of certain pieces or sets of data, which would make them informationally significant in this economic sense.

Each of the three types of information can further be understood as a) information as its own entity (for example, bits in a computer memory), b) information about something (for example, a timetable about train departure times), or c) information abstracted from something (for example, patterns identified in wasp behaviour). Accordingly, a thought experiment could be studied as a) containing information, b) providing information, or c) doing something from which we can make useful judgments about the information in/about something else.

How might a thought experiment play these roles? According to the semantic characterization of information, thought experiments could increase information by venturing away from what is known. We can concoct imaginary scenarios that are less predictable given what we know than actual scenarios, which by this sense of semantic information would be more informative, as long as they are truthful, meaningful and well-formed. Consider Derek Parfit’s thought experiment about people splitting like amoebas (1986, 254). In this thought experiment, Parfit asks what we would say about a person who split into two organisms, while remaining conscious the whole time. At the end of the splitting process, the original person is psychologically
continuous (= identical?) with each of the people after the split, but those people are not psychologically continuous with each other. And there cannot be two (different) creatures which are also the same creature, because this violates the transitivity of identity. This thought experiment has been criticized for being too outlandish (Wilkes 1988, 36), but on this characterization of empirical import, it is more informative than considering actual cases. Interestingly, this reply to Wilkes is very much in line with Parfit’s actual response to the problem of using outlandish examples, although he does not use the language of information or informativeness (Parfit 1984, 255; and more generally on 200). The debate between Parfit and Wilkes might therefore be reconceived as a debate over how much truthfulness is required for something to count as semantic information.

As above, one way to solve the paradox of thought experiments characterized in terms of information is to argue for a certain philosophical account of information, in the hope that such an account will explain how empirical information can be increased by the mind without the need for new empirical data. And just as before, the tables can be turned; we can also consider specific thought experiments as a way of testing different philosophical accounts of information.

While no one has tried to explicate the paradox of thought experiments in terms of information, some, like Parfit, have hinted at it. It is an interesting angle from which to view the problem, since the resources in the philosophy of information would allow us to explore changes in epistemic information via thought experiments in a formal way.

3.2.3 Empirical Import = Empirical Evidence

Evidence has been characterized in many different ways. It could be whatever provides justification for a proposition (Kim 1988), the set of all one’s knowledge (Williamson 2000), sense-data (Russell 1912[1997]), observation statements (Quine 1968), the set of all one’s occurrent thoughts (Conee and Feldman 2004), a phenomenon (Brown 1993b), or it could be a physical thing like a murder weapon. And of course many of these characterizations overlap. Under some of them thought experiments are evidential, and under others they are not. Specifically, thought experiments do not provide sense-data, although they can present and manipulate it in interesting ways. They also do not provide observation statements, unless we count introspection or intuition as observation. And they do not present us with physical objects
like murder weapons. While they do not provide “all one’s knowledge,” they might provide a new piece of knowledge. The same goes for the set of all one’s occurrent thoughts, since new thoughts are certainly possible through thought experiments. Also, many have argued that thought experiments can provide justification for a proposition, theory, etc. And Brown argues that they can introduce new phenomena. There are therefore many ways to understand evidence according to which thought experiments are evidential.

To fit the form of the paradox, the evidence must be empirical. What does it mean for evidence to be empirical? For one, it could be based on sense experience. Prima facie we might think that evidence provided by thought experiment is not based on such experience, and indeed, thinking otherwise contradicts the many authors who claim that thought experiments are a priori, or in other words, independent of experience. But this apparent contradiction can be resolved if some of the material that thought experiments draw from is based on experience. A thought experiment about what is happening right now in the town hall of your nearest neighbouring city will not be based on experience in the sense of providing direct observation, but the memories and representations that figure into that thought experiment will be based on sense experience. The “based-on” relation seems to be transitive. Thought experimental evidence based on inferences, which are based on memories, which are based on experience, is still in some sense based on experience. The sense in which thought experiments are a priori then, if they are, concerns the source of the justification for the conclusion, not the source of the content of the mental items that feature into the thought experiment.

Another way to characterize the status of evidence as empirical is to highlight the fact that empirical evidence is verifiable (or verified) by sense experience. According to this characterization, many thought experiments present evidence that can be made empirical by attempting to confirm it empirically. Galileo’s falling body thought experiment is made empirical when astronauts drop a hammer and a feather on the moon. An objection is that many thought experiments are not so confirmable, for example, because they rely on frictionless planes, demons, or situations where we must decide whether we should kill one person to save five, etc. Some philosophers have nevertheless argued that this is a good characterization of empirical evidence; one that good thought experiments should strive to meet. Buzzoni for instance, claims that all thought experiments should be empirically verifiable in principle.
(Buzzoni 2010, 2013a. Also see below, Chapter 8 section 3.1. For criticism see Fehige 2012, 2013. For a reply see Buzzoni 2013b).

As above, we might attempt to solve the paradox by adopting one of the several existing epistemological strategies that explain how justification of empirical belief takes place. That is, we can be rationalists, empiricists, naturalists, and so on. And again, there is the option to turn the debate around and use the study of thought experiments to tell us something about the feasibility of each of these epistemological positions. Both directions of argument have been employed by philosophers, as we will see below.

3.3 Inside the Head

Finally, what do we mean when we say that thought experiments are conducted “entirely inside one’s head.” Here are three possibilities: a) the source of the elements manipulated in the thought experiment are in the head, b) the location of the thought experiment itself is in the head, or c) the source of the evidence that justifies the output of the thought experiment is in the head.

Fig. 3: How a Thought Experiment Might be “Inside the Head”
We might still consider a thought experiment that relies on a diagram or the waving of one’s hands to be entirely inside the head, despite its reliance on something external, if that reliance is not justificatory as in c). For example, Stevin’s thought experiment about a chain draped over a frictionless prism is almost always accompanied by a diagram. But this diagram is not what justifies the conclusion of the thought experiment, although it is helpful in reaching that conclusion. So it is inside the head in the sense of c), although perhaps not a).

On the other hand, there are thought experiments which might rely for their justification on extra-mental features. What makes a consideration extra-mental in terms of justification is of course hotly debated. For some German idealists and Berkeley, very little counts as truly extra-mental. Other philosophers will draw the line in different places.

Locating the mechanism of the thought experiment itself provides even more options. Since imagination is involved in most thought experiments, and imagination is not entirely within one’s head but often related to kinaesthetic senses, then perhaps no thought experiment is internal in the sense of b). “If you imagine, for instance, lifting a heavy weight, there will be electrical activity in the muscles in your arm, even though your arm does not actually move. Probably, the activity in the muscles (and the signals the muscles send back to the brain) is just as much a part of the imagining as is the activity in the brain that signalled the muscles to move, but then told them not to move after all” (Thomas 2011). If our minds include our nervous and muscular systems, then perhaps the paradox should be augmented to reflect this. But the literature on embodied cognition does not stop at our fingers and toes; some have argued for the more ambitious “extended mind thesis,” which allows items in our environments (like personal notebooks) to count as part of the mind (see for example, Clark and Chalmers 1998). Others extend even further to include the minds of others (for example, Longino 1990). Wherever we draw the line between inside and outside, the point seems to be that thought experiments are carried out in such a way that the target system is not investigated by the senses, but rather the mind (whatever that is).

Concerning a), note that the location of the sources on which a thought experiment draws depends on what the sources are, and how they relate to the mind. In some sense, the mind certainly “captures” features of natural systems, just like a computer model or laboratory
experiment does. But how? Was the object of inquiry (for example, KNOWLEDGE, GOODNESS, SIMULTANEITY) already in our minds? Perhaps the external object remains “outside,” while part of it is abstracted and is “taken inside,” for example, a few of its properties. Or perhaps some features of the system are re-created fresh as mental representations, as an artist recreates with a portrait, or an engineer a wind tunnel.

As above, we could try to explain the way that manipulations of ideas that are “entirely” inside the head can tell us something about the world outside it, or we can use thought experiments to tell us something about what can happen inside the head, or where the line should be drawn between epistemologically internal and external.

The conceptual space of the paradox can therefore be represented like this:

Fig. 4: (Some of) the Conceptual Space of the Paradox

For each account in the literature, there is a way that it interprets the paradox of thought experiments. And in each case, there is an alignment of these wheels that corresponds to that

---

3 Throughout this dissertation I contrast thought experiments with laboratory experiments. Other authors use “empirical,” “material,” “real,” or “actual” in place of “laboratory.” I prefer “laboratory” for the following reasons. Not all extra-mental experiments are empirical in the sense of relying for their justification mainly on sense-experience. And not all of them are material in the sense of investigating only or primarily the material properties of a system. And I think thought experiments are both real and actual. Therefore while I recognize that not all experiments are performed in a laboratory, I use that term to contrast thought experiments with the extra-mental, contextual and historical practice of publicly accessible experimentation.
interpretation. Some of the lines on the wheels should be blurry, and I only considered what I thought were the most attractive options. There are still many other possibilities. For each additional interpretation of “novelty,” “empirical import,” and “inside the head,” the number of possible accounts is multiplied. Still, for every alignment of the wheels, there is a version of the paradox to be answered.

Accounts in the literature on thought experiments are in danger of talking past one another if they address different interpretations of the paradox. Explaining how thought experiments can produce something surprising about the psychological empirical content of a concept while only relying on memory is very different from explaining how thought experiments can provide brand new experiences which justify empirical knowledge relying only on the light of pure reason.

4 Types of Account: Cartesian and Baconian

An interesting pattern emerged while mapping the different interpretations of the paradox. When we considered empirical import as empirical semantic content, positions in the philosophy of language became relevant. When we considered empirical import as empirical evidence, positions in epistemology suggested themselves. In each case, we saw that positions in general philosophy could be employed to resolve an interpretation of the paradox, or conversely, facts about thought experiments could be used to tell us something about those positions. For example, Norton uses empiricism to answer the question of how our minds can produce novel empirical evidence through thought experiments. Conversely, Brown uses specific thought experiments to try to prove empiricism inadequate. I will affectionately label the first direction of application “Cartesian” and the second “Baconian.” René Descartes, as is well known, believed that from a clear and distinct set of philosophical principles we could derive all knowable truths about the physical world. The Cartesian approach to the paradox of thought experiments therefore tries to derive the surprising and wonderful facts about thought experiments from a set of philosophical assumptions about meaning, information, evidence, epistemology and the mind. John Norton is Cartesian in this sense as he tries to show how the assumptions of empiricism can account for the seemingly miraculous things thought experiments can do.
Francis Bacon, on the other hand, was more accustomed to messy courts of law than to the beautiful deductions of mathematics. He argued that we should give up trying to fashion deductive systems that predict all possible observations with certainty, and instead stay as close to the facts as possible. Do not create and compare theories built to encompass the observations before all possible observations are made, he pleaded. Instead, gather all the observations you can, no matter how confusing they may seem. Afterwards, find an explanation that provides power over the phenomena in question. Then make this explanation coherent with other power-giving explanations. In the end, we will be left with an explanation that explains all of our explanations, and therewith, our experience. Bacon stressed that this final explanation should be the crowning achievement of science and not a foundational assumption. An example of this strategy as applied to thought experiments is Brown. While Brown subscribes to Platonism, he takes the features of thought experiments to justify adopting this stance, and not the other way around.

Now, let us see how the main accounts in the literature understand the paradox in some of the different ways listed above, and how they fit into the meta-level categorization of strategy-types into Cartesian or Baconian.

5 Above Distinctions Applied to the Modern Epistemological Accounts of Thought Experiments

Brown takes the novelty of thought experiments very seriously. For example, Brown claims that some thought experiments can bring us to “see” something new. Although he is speaking metaphorically, he and other rationalists argue that the mind can grasp abstract objects or the relations among them. Those who claim that thought experiments can be a priori, including Brown (2011) and Hopp (2014), posit what I think is the strongest sense of novelty. For these authors, what is novel is an experience that takes the form of a mental perception. It is novel because we have not had it before, or if we have, it now takes on a new importance because it answers a question we were not previously posing. And because this new experience fallibly justifies an inference, it provides a change in the epistemological status of a belief.

Thus for Brown, Galileo’s falling body thought experiment provides novelty in at least six senses. 1) The thought experiment is surprising, 2) its conclusion is not derivable, 3) it provides a
new experience (a mental perception), 4) that experience demonstrates a new relation between beliefs (concerning universals), 5) this relation is a new law of nature (the speed of free-fall does not depend on the weight of the object) that we now believe, and 6) we have new justification for that belief (rational intuition).

How should we understand the epistemic import of a thought experiment for Brown? Destructive thought experiments (1991, 33) are reductio ad absurdum arguments or counter-examples to theoretical claims, which can be understood as providing empirical evidence. Constructive and Platonic thought experiments establish well-articulated theories (1991, 40), so they also produce evidence. In Platonic thought experiments we are given a glimpse of the laws of nature, which is empirical because those laws delimit the modal landscape for the empirical world they govern. The evidence is empirical, therefore, because it is relevant for experience, not because it is based on sense experience. Other thought experiments illustrate or make plausible some theoretical claim (1991, 35-36), which could be a weaker form of evidence. If these thought experiments provided no evidence at all but merely helped us to work out what was going on in a theory, then it would have empirical import in the sense of empirical psychological content. Therefore Brown incorporates at least two of the three main interpretations of empirical import.

Finally, the imagined scenarios in a thought experiment are “in the head” for Brown. The same goes for the process of thought experimenting itself. But the abstract entities related in laws of nature and mathematics are mind-independent, and exist in some sense outside of the head. This adds an interesting caveat to Brown’s version of the paradox concerning Platonic thought experiments, which really asks: how do thought experiments produce (genuinely) novel empirical evidence (in the sense of experience that justifies belief) without new empirical data, allowing for new non-empirical data?

Norton claims a different sense of novelty for thought experimental conclusions than Brown. For Norton, thought experiments are arguments that “draw from what we already know” either tacitly or explicitly, and then “transform” that knowledge by some form of deductive or inductive inference (2004b, 45). Since Norton is an empiricist, his “account of thought experiments is based on the presumption that pure thought cannot conjure up knowledge, aside, perhaps, from logical truths. All pure thought can do is transform what we already know” (49).
Since deductive arguments can only rearrange existing information, what emerges might surprise us, but it will not really be new. Norton writes, “Deductive inferences merely restate what we have already presumed or learned. If we know all winters are snowy, it follows deductively that some winters are snowy. There is no mystery in what permits the conclusion. We are just restating what we already know” (Norton forthcoming, Chapter 2). Deductive thought experiments can produce psychologically novel outcomes that change the epistemological status of a belief, but they are not going to provide something that was not derivable from the premises of the thought experiment and its background theory, and whatever psychological experiences they cause will not be epistemologically relevant.

Inductive arguments produce a little more novelty. Norton writes, “I shall use ‘induction’ and ‘inductive inference’ as the general term for any sort of ampliative inference; that is, any licit inferences that lead to conclusions logically stronger than the premises” (Norton forthcoming, Chapter 1). So besides new beliefs and justification for those beliefs, thought experiments can also expand the logical scope of a proposition when it is inductive.

Turing to empirical import, while Brown is concerned with the panoply of ways that thought experiments can increase human understanding, for example, by illumination, explanation, theory-change, etc., Norton is concerned with the following question only: “Thought experiments are supposed to give us knowledge of the natural world. From where does this knowledge come?” (2004b, 44). Norton therefore understands empirical import as empirical evidence for knowledge claims. On Norton’s characterization of the paradox we must therefore explain how thought experiments can justify beliefs that are true and surprising. These beliefs will be empirical in the sense that they receive their content from previous sense-experience, but also in the sense that they are relevant for experience.

The majority of philosophers place themselves somewhere between Brown and Norton. I will start with those closer to the Norton-end of the scale and move towards Brown. Sören Häggqvist (1996, 2009) agrees with Norton that insofar as thought experiments play an evidential role in science or philosophy, they must play a part in an argument. That is, thought experiments are used to contest or bolster theoretical claims by providing (usually modal) evidence that counts for or against the claim. For Häggqvist, the thought experiment plays a justificatory role in the
same way that a real experiment does: by contradicting or supporting a claim made by the theory. This makes them (parts of) arguments, but only in a general sense. Häggqvist denies that the performance of a thought experiment is the performance of an argument, for one, because a thought experiment cannot be formally valid or invalid. Häggqvist’s insight is taken up by Tim De Mey (2003), who argues that we should investigate the epistemic impact of the thought experiment’s conclusion in one way, and the nature of the thought experiment itself, in another.

If Häggqvist and De Mey are right, this gives us an interesting way to combine two senses of empirical import for thought experiments. First there is the conclusion of the thought experiment, which is a product of psychological mechanisms and is somehow related to experience. Then there is the use of that conclusion, however justified, in an argument for or against the truth of a claim about the world. If thought experiments are to provide items of “novel” empirical import, they might do so in either or both of these ways, and each would be subject to different epistemological explanations. For example, the mechanisms which justify the production of the thought experimental conclusion might involve faculties of sense perception, imagination, memory and intuition, which can be tested for reliability and accuracy. The mechanisms which justify the use of a thought experimental conclusion in the production of new theoretical knowledge might be the ones identified by logicians, like modus ponens and inference to the best explanation.

This brings us to the naturalists, that is, to those who claim that we can and should turn to science to discover how thought experiments work. This idea is clearly present in Ernst Mach (1905) and Wolfgang Yourgrau (1962, 1967). But Roy Sorensen was the first to give an in-depth expression of naturalism about thought experiments (1992).

Like Häggqvist, Sorensen sees thought experiments as a type of modal reasoning. And like Mach (1905), he places thought experiments on a continuum with real experiments. Along with several others (including Lichtenberg, Kuhn, Gendler, Bokulich and Arthur), Sorensen argues that thought experiments mostly eliminate irrationalities in our thought. And again following Mach, he claims that thought experiments function by drawing upon the stores of empirical knowledge that we personally accumulate combined with the innate ideas and structures that have been programmed into our minds by evolution. Sorensen stays in-line with Norton, arguing that
thought experiments mostly “repackage” old information in a way that makes it “more informative” (1992, 4). But his interpretation of empirical import is quite different from Norton’s. Instead of empirical knowledge, he takes the goal of thought experiments to be the creation and stabilization of phenomena, atheoretical exploration and the definition of concepts. Achieving these goals does not amount to creating new empirical knowledge in the sense of providing new justified beliefs. Creating phenomena in the mind is not knowledge without some accompanying experience telling us that what we have created is accurate or useful in understanding some phenomena. Second, assuming that there are atheoretical concepts to explore in science, such exploration might be a necessary aspect of science, but it is an aspect that an empiricist would relegate to the context of scientific discovery and not justification. Exploration is not knowledge. Finally, defining a concept creates analytic or logical truth, and not empirical knowledge. Sorensen is therefore much closer to Brown than he is to Norton with respect to the empirical import of thought experiments.

Since Sorensen, many more naturalists have emerged. A subset of these focus on the performative aspect of thought experiments, and characterize them as “mental models.” In so doing they bring us still further from Norton. “Mental model” is a technical term of cognitive science (see Johnson-Laird 1983) and was first applied to thought experiments independently and simultaneously by Nenad Miščević (1992, 2004, 2007) and Nancy Nersessian (1992, 1993, 2007, 2008), who agree that thought experiments are used to mobilize special skills of the experimenter which we might vaguely characterize as knowledge how.

Nersessian argues that the narrative presentation of a thought experiment triggers the creation of a “discourse model” which is a “representation of the spatial, temporal, and causal relationships among the events and entities of the narrative” (1992, 294). Such a “mental model” is often visual in nature, and is manipulated in real time. It draws on embodied wisdom (294) and embeds a specific and personal point of view into the model (295). With respect to the paradox, Nersessian makes a telling remark. “The constructed situation, itself, is apprehended as pertinent to the real world in several ways. It can reveal something in our experience that we did not see the import of before…It can generate new data from the limiting case…[and] it can make us see the empirical consequences of something in our existing conceptions” (296). In other words, Nersessian recognizes some of the different ways empirical import can be interpreted. Instead of
producing new knowledge, a thought experiment can highlight old data that did not seem important, it can separate phenomena that seemed necessarily connected, it can generate new data from limiting cases, and it can make clear the consequences of previous conceptual commitments.

Miščević (2007) agrees that “the mental model proposal can account for the justification of intuitional judgements within a more naturalist framework than the one endorsed by Brown” (182). Miščević recapitulates the phenomenology of (visual) thought experiments by distinguishing the stages through which they generally progress. Roughly, what happens is that “in the first stage the cognizer tries to imagine a scene verifying a given proposition. Sometimes a testing stage follows, in which one tries to imagine situations that would falsify some given proposition. In the last, recapitulating stage, the cognizer typically first judges that no imagined situation falsifies the proposition, second, assumes that she has inspected all imaginable situations, and third, infers that it is impossible that the proposition tested is false, i.e., that it is necessarily true” (194). Miščević recognizes that the “it is necessarily true” step is not always justified, but he adds it because descriptively speaking, we do often take this step.

He also argues that many problems are easier to solve when represented in a mental model as opposed to verbally or formally. This is because you get to use the same faculties you use to understand everyday real-life situations (like seeing things from different perspectives, or moving in a gravitational field), and this mobilizes tacit knowledge. Our familiarity with the features of everyday scenarios explains the rapidity with which we see what happens in an imaginary scenario. This is unlike having to work out a solution deductively, which presumably draws on less-practiced skills.

Tamar Gendler is another proponent of the mental model view. She asks how “contemplation of an imaginary scenario can lead to new knowledge about contingent features of the natural world” (2004, 1152). This means she interprets the paradox like Norton, where empirical import means empirical knowledge. Her solution is that in a thought experiment we consider imaginary scenarios which evoke quasi-sensory intuitions, and this process can lead us to new beliefs, which are justified if they are produced by a sufficiently reliable cognitive process, which she argues they are.
Gendler claims that Norton must be wrong, however, because the phenomenology of thought experiments simply will not allow that thought experiments can be arguments. Quasi-perceptual, imaginative reasoning is not argumentative reasoning. She gives three types of counterexample: manipulating mental models, using the imagination to trigger emotional responses, and changing perspective to come to some new justified beliefs. Not all thought experiments require these actions, but since there are some that do, she concludes that Norton cannot be correct about thought experiments in general.

What can we say about the sense of novelty invoked by those who characterize thought experiments as mental models? Gendler, Nersessian and Miščević all agree with Norton that existing knowledge can be manipulated and transformed in a thought experiment to produce something novel. For example, a thought experiment can restructure old information in a new way that reveals features of that information that were not previously noticed. While the structure is new, the content is not, so this might fall under the psychological sense of novelty.

However they claim more novelty than Norton by adding other senses of the term that go beyond mere rearrangement and change of scope, yet less than Brown. For example, rearranging old data in new ways can highlight salient features of that data. Drawing attention to features that were not noticed before approaches the sense of novelty in Brown because it considers novel mental “presentations.” But it does not reach Brown, because the presentations are not interactions with real, mind-independent abstract entities. Gendler, Nersessian and Miščević agree that rearrangement can produce new concepts or help us to “possess” old ones by making them our own, as when a thought experiment helps us to overcome a fear of flying that we know is irrational. Statistical knowledge that flying is safe is not enough to prevent fear in some people, yet thought experiments in the form of repeated positive visualizations can help make their statistical knowledge about the safety of airline travel useful (Gendler 2004, 1160). This sense of novelty concerns our abilities, and the relationships between our beliefs.

For these authors thought experiments can also produce genuinely new concepts, and when they do this, they produce something that was not derivable from the propositions given in the thought experimental set up.
Finally, it is sometimes hinted that what provides the novel empirical import in a thought experiment is the exercise of a modal faculty of intuition that is stimulated by the thought experiment. For example, Ichikawa and Jarvis (2009) argue that it is not just background assumptions that we rely on in a thought experiment, but our ability to interact with *stories*. Perhaps the human brain has evolved some reliable way of forming modal inferences from imagining what would be true in a fictional world, and thought experiments take advantage of this. Of course, for this to be a new account it must be argued that the manipulation of imagined scenarios and the filling-in of those scenarios by the use of tacit knowledge are *not* merely acts of argumentation. Talk about modal faculties must not be elliptical for talk about counterfactual arguments.

Now, are philosophers who characterize thought experiments as mental models Cartesians or Baconians? Insofar as they are committed to a certain view of psychology which demands that they understand human reasoning a certain way, they are Cartesians who approach the paradox of thought experiments with the tool kit of the naturalist, ready to synthesize what they find. This is why the conflict between the mental modellers and Norton is such an interesting one: each side thinks their starting point is correct and capable of explaining all the defining features of thought experiments (Norton 2004b, 60-61).

However the mental modellers also seem Baconian, as there is enough scientific evidence, they argue, to believe that humans really do reason in terms of mental models. And this is what justifies their Cartesianism, not a set of foundational philosophical assumptions. However, their argument relies on the philosophical assumption of naturalism, which of course cannot be justified by Baconian induction.

Complications like these are bound to arise when adopting this somewhat naïve characterization of the strategies in the literature. The distinction between Cartesian and Baconian strategies should not be taken too seriously. Still, despite its inexactness, I think it can be used to explain some features about the Brown-Norton debate.

For one, the distinction explains why the *initial* response of students and philosophers is to side with Norton. This is because given a choice between two Cartesian approaches, one from the assumptions of empiricism and the other from the assumptions of Platonism, most people will
pick empiricism. However Brown does not employ the Cartesian strategy, and immersion in the literature makes this clear. Since I think most philosophers of science look to apply the Baconian strategy, they side with Brown on his meta-theoretical approach. This is why there are more philosophical articles that criticize Norton’s account than Brown’s: philosophers have taken issue both with the specifics of Norton’s account and his Cartesianism.

Another very interesting thing to notice is that almost everyone who attacks Norton does so in a Baconian way. For example, Brown has told me in conversation that after first encountering Norton’s arguments, he went about finding counter-examples that would disprove the thesis. This is a common reaction, and several papers have been published which claim to find such cases (for example, Bishop 1999, Brown 2007, Gendler 1998). These papers are trying to fight a Cartesian using Baconianism, that is, they are looking for observations to disprove a fundamental theory. These are unlikely to convince Norton, however. What observation could convince Descartes that motion was not vortical? Unless you disprove empiricism itself or show Norton’s account to be incoherent, it is unlikely that anyone will convince Norton by examples.

6 Chapter Summaries

Since it is rare to see someone take up Norton’s account on its Cartesian terms, this is what I do in Chapters 2 and 3. I ask in Chapter 2 whether Norton’s attempt to solve the paradox works when we grant his assumptions. In Chapter 3, I ask whether we should grant those assumptions.

Specifically, in Chapter 2 I ask what Norton means by “argument.” He allows thought experiments to be inductive, deductive, abductive and informal arguments. But when examined more closely, it turns out that his characterization must be so broad that it allows any inference that is convincing to be an argument. And this makes his position trivial, given that thought experiments are convincing.

In Chapter 3, I ask what justifies a thought experiment, according to Norton. His answer is that a thought experiment is justified when it displays a “mark” that can be identified in the logic of the inference connecting premises to conclusion, where that mark is whatever present or future logic identify as reliable. I argue that we should not evaluate thought experiments (or arguments) this
way. What makes a thought experiment good or bad is not a logical categorization; those categorizations only group together inferences we already approve of for other reasons.

After finding Norton’s approach unsatisfactory, I return to the Baconian approach, which has recently been faced with an important challenge. From the facts about thought experiments, we collectively infer that there is a paradox. The cogency of this judgment has been called into question by Paul Thagard (2010a, 2014). Thagard claims that we have not been careful: if we look closely at thought experiments, they do not support this inference. They are not reliable. They do not provide novel empirical import. The very fact that they operate “entirely” from inside our heads ensures this. There is no paradox.

Chapter 4 argues that Thagard is attacking an idiosyncratic version of the paradox: a version that claims thought experiments can provide a special kind of empirical evidence (necessary a priori knowledge). I refute this attack. My refutation is not a refutation of all skeptical challenges to thought experiments, since as we have just seen, there are many other ways of forming the paradox, and Thagard’s skeptical challenge only addresses one of them (for others, see Dancy 1985, Dennett 1984, Duhem 1954, Harman 1986, Meinong 1907, Wilkes 1988). Thagard rightly accuses philosophers of overlooking many failed, misleading and dangerous thought experiments. While I reject Thagard’s interpretation of the paradox, I recognize that we must secure a version of the paradox that is answerable.

This becomes my project starting in Chapter 5. I argue that an examination of historical case studies inspires an answerable version of the paradox. After I look at the some important scientific thought experiments with a focus on the relationship between thought experiments and theory, I suggest an interpretation of the role of thought experiments that I think is new. I claim that thought experiments increase understanding in science by increasing the empirical content of theoretical structures. By “theoretical structure” I mean the concepts, models, theories and principles that make up the practice of science. I use the word “structure” to emphasize the conceptual connections between the elements of the structure and other structures. $F = ma$ is a structure in this sense. But I do not want to imply that they are structural as opposed to material. That is, I am not claiming they are devoid of content. I would also like to emphasize that how much and what content a structure has is relative to the individual. This position does not lead to
relativism because there is no ideal empirical content for any theoretical structure, just like there is no ideal content for GOLDEN RETRIEVER. Being able to point out a golden retriever or recognize one from a description is enough empirical content for basic competency, but not for a dog show judge—and it is the same in science. When a theoretical structure is first devised by a scientist or encountered by a student, it is often the case that the empirical content must be increased for that individual before they can employ it well (and like any concept, its content continues to evolve over time). I argue that many thought experiments in science can be used to provide novel empirical content in this sense. Such thought experiments are successful when they create new abilities, such as being able to use the theoretical structure to interact in new ways with experience, other people, and other structures. I argue that the resulting epistemically desirable state is understanding (as opposed to knowledge), and I reject the need to locate thought experiments inside one’s head by aligning myself with embodied cognition.

These conclusions are the result of case studies undertaken in Chapter 5, so I do not claim my conclusions to be applicable to all thought experiments. For example, many mathematical thought experiments might function in a similar way, however they would not be increasing empirical content but rather some other kind of content (rational, perhaps). My conclusions are therefore only interesting if the set of thought experiments I analyze is.

In Chapter 6, I test my interpretation of historical thought experiments in science against studies of subjects performing and learning from thought experiments in real time. These results suggest an epistemological account that I develop in Chapter 7.

Here are some of the conclusions of the final chapters. 1) There are thought experiments that increase empirical content, where empirical content is semantic content that is empirically relevant. 2) Part of the output of such thought experiments should be characterized as increasing understanding. 3) A thought experiment can enrich scientific understanding while simultaneously being useful for criticizing a theory, encouraging conceptual change, justifying beliefs, serving as evidence, etc. That is, a thought experiment can increase knowledge while also increasing understanding. 4) The thought experiments I analyze increase empirical content by helping us to connect theoretical structures to other theoretical structures, experiences, abilities, emotions, or values, via an exercise of the imagination. 5) The imagination tries out
different connections, one of which is psychologically “promoted” (in a fallible way) for its intelligibility and potential fruitfulness. 6) Knowledge and understanding are closely related, but independent. In some contexts, possession of knowledge can necessarily imply the possession of some associated understanding.

Here is a schema for the thesis in terms of the paradox:

Introduce and Analyze the Paradox  
\[ \rightarrow \]
Adopt Norton’s Characterization of the Paradox  
Show Norton’s Account Unsatisfactory  
\[ \rightarrow \]
Adopt Thagard’s Characterization of the Paradox  
Show Thagard’s Account Unsatisfactory  
\[ \rightarrow \]
Identify a New Characterization of the Paradox by Historical Case Study  
\[ \rightarrow \]
Explore the New Characterization using Social and Cognitive Science  
\[ \rightarrow \]
Sketch a Preliminary Account to Explain how Thought Experiments Play the Role Identified by the New Characterization (That is, How Thought Experiments Produce Scientific Understanding)  

One final theme that connects the chapters of this thesis is experimentation. Norton’s view denies the experimental character of thought experiments, and Thagard overlooks it. Descriptively, I argue that empirically relevant thought experiments (or arguments) will always depend on semi-experimental interaction with experience. Chapter 4 argues that thought experiments can profitably be portrayed as mental models, which allows me to focus on their experimental features by highlighting the similarities between model-based reasoning and other semi-experimental methods such as computer modelling. Normatively, this implies that the
criteria for good laboratory experiments will be just as important as logical considerations are for evaluating thought experiments. Finally in Chapters 5-7, I provide an account that grounds the power of thought experiments to produce scientific understanding in terms of experimental connections tried out by the agent in his or her imagination.
Chapter 2
Thought Experiments and Arguments

Invention is not the product of logical thought, even though the final product is tied to a logical structure.

Albert Einstein (Pais 1982, 132)

The notion that thought experiments are nothing more than arguments in all epistemologically relevant aspects is an influential view in the philosophy of thought experiments. John D. Norton is its principle standard-bearer (Norton 1991, 1996, 2004a, 2004b). Following Brown and Fehige (2014) I refer to this position as the “argument view.” Its refutation is the goal of this and the next chapter.

Norton claims that his view is a consequence of empiricism (1996, 334). This is significant for Norton because empiricism is “overwhelmingly the predominant epistemology in philosophy of science” (2004b, 50), and therefore the argument view should be the default position in the philosophy of thought experiments.

To contest Norton’s account, one might deny that empiricism is (or should be) the predominant epistemology in the philosophy of science, or that the argument view follows from empiricism. Instead, I will adopt the empiricist’s assumptions, allow the inference from empiricism to the argument view, and argue nonetheless that the argument view is circular (this chapter) and incomplete in a philosophically worrying sense (next chapter).

Norton wants to dispel the apparently rationalist flavour of thought experiments, which comes as a result of the fact that they appear to yield information that is empirically relevant without the need for new empirical input. Norton claims that the only non-miraculous way to get new information about the mind-independent world without empirical investigation is by performing logical manipulations on existing knowledge. If thought experiments produce new information about the world it is because they are analytic tools of reason that merely bring to light the
hidden consequences of and relations between facts. Thus, if we wish to eliminate appeal to “epistemic magic” (2004b, 45), we must accept that thought experiments are arguments.

Norton provides two other reasons to think the argument view is correct. The first is the “reconstruction thesis,” which asserts that all thought experiments can in fact be reconstructed as arguments. This is meant to be a surprising piece of information that demands explanation, which the argument view provides. In other words, the best explanation for the reconstructibility of every thought experiment is that thought experiments and arguments really are the same thing.

The other support for Norton’s argument view is the “reliability thesis,” which Norton claims is “independent of empiricism” (Norton 2004b, 52). The reliability thesis asserts another surprising fact: thought experiments are always equally justified as the arguments into which we reconstruct them. Arguments are the only cognitive tools that command the power of logic, and logic governs all successful inferences. So if thought experiments are epistemically reliable cognitive tools then it makes sense that they would be arguments.

In this chapter, I present the details of the argument view as a way of extracting and analyzing the notion of “argument” to which Norton is committed. What does Norton mean and what should he mean by “argument”? Such an analysis has never been attempted, and I think it is necessary for any meaningful critique of the argument view. In the course of this analysis, I show that the identification of thought experiments and arguments is circular if arguments are the kinds of things that Norton is committed to saying they are. This chapter is therefore an attack on the reconstruction thesis. But unlike other attacks, which try to point out some feature of thought experiments that cannot be reconstructed as part of an argument (for example, Bishop 1999, Brown 2007a, Gendler 1998), this one questions the coherence and desirability of reducing thought experiments to arguments, as Norton conceives of them.

1 Some Context

Despite the attractive simplicity and austere ontology of the argument view, it has dissatisfied many philosophers writing about thought experiments (for example, Brown 2011; Cooper 2005; Davies 2007; Gendler 1998; Gooding 1994, Häggqvist 2009, Hopp 2014, Miščević 2007; Nersessian 1992, 2007). The widespread uneasiness stems, I contend, not from the assumption of
empiricism, but from the supposed inability of arguments to perform all the functions we assign to thought experiments. Usually, when the argument view is addressed, some specific feature of thought experiments is identified that apparently cannot be captured when we reconstruct it as an argument. For example, James R. Brown says that thought experiments create a kind of mental phenomena, which arguments cannot do if they operate solely with propositions (2007a).

Michael Bishop points out a historical case in which the same thought experiment is used to draw two opposing conclusions, something an argument cannot do (1999). The reason this type of complaint will not refute Norton’s reconstruction thesis is due to the flexibility of his notion of argument.

No one has tried to extract Norton’s conception of argument by a comprehensive and careful reading of his work on the subject, although Norton seems to recommend such a project himself. In correcting a “persistent confusion,” he writes:

Some (for example Gooding 1992, 283; Hacking 1992, 303) report that I demand the argument in a thought experiment must be deductive; others suggest the argument must be symbolic (or so I have seen reported in a manuscript version of a paper); and others (Boorsboom et al., 2002) that the arguments must be derivations within some definite theory. A brief review of what I have written will show that none of these restrictions are a part of my view, which allows inductive and informal argumentation and premises from outside any fixed theory. (Norton 2004b, 64)

Given that there is this confusion about how to interpret Norton’s view, this chapter will consider his conception of argument from the earliest papers to the most recent. It is surprising that no such study has been done, because in order for Norton’s “deflationary view” (2004a, 1140) to succeed in dissolving the epistemological problem of thought experiments by appealing to empiricism, he must recast it as a problem that has already been solved by the empiricist epistemology of arguments. (“Thought experiments are epistemologically unremarkable” Norton 1996, 334). And to do this, he must employ a consistent notion of argument that is acceptable to his opponents (or at least to empiricists), which fits together with the reconstruction and reliability theses.
2 The Kinds of Inferences that Count as Arguments

Here is Norton’s earliest statement featuring the argument forms with which thought experiments can be identified.

A very broad range of argument forms should be allowed here; in particular they should include inductive argument forms. In the thought experiments of modern physics, the premisses of the arguments are generally held within one or other physical theory. Thus a thought experiment might be a purely deductive derivation within some physical theory. But not all such thought experiments are. As we shall see later, some do involve inductive reasoning or the introduction as a premiss of a general philosophical principle not contained within any specific physical theory. (1991, 129-30)

In this passage Norton identifies thought experiments with inductive and deductive arguments. Deductive arguments for Norton are those that “reorganize” assumptions, and inductive arguments are those that “generalize” them (1996, 335).

Norton (1996) expands his conception of argument to include inference to the best explanation (1996, 349). He also adds informal arguments in the following way. After summarizing and reconstructing a mathematical picture proof presented by Brown that provides evidence using a diagram (Brown 1993a, 275-76), Norton concludes,

It is clear from the way the commentary is laid out that I think it is an argument. Each labelled line is a step. The unlabeled lines are padding. The ‘from figure’ steps are just another form of premise/assumption. Now this is clearly not a proof in the formal sense, but it is an informal argument good enough to convince all but the sternest skeptics…The thought experiment is convincing; the corresponding argument is informal but strong. (1996, 353)

Norton counts Brown’s picture proof as an informal argument, and he later says that he has “always urged that thought experiments may be informal arguments” (2004b, 58). It is very interesting that Norton allows diagrams to be used in informal arguments. If the ‘from figure’ steps are just another form of premise/assumption, then a diagram or picture can replace a
proposition (or element of a proposition). This is so because for most empiricists, it is propositions that make up arguments. In this case, Norton is committed to the possibility of diagrams being true or false, because propositions that figure in arguments are capable of being true or false (or admit some probability of truth). This is a substantial and controversial view (see Baird 2004, French 2003, Meynell 2009, and Perini 2005). If arguments work with propositions, and if diagrams can replace propositions, then an activity like reading a map or a patient’s x-rays can be a step in an argument, and those maps and x-rays (or our gazing at them) must be capable of being true, false, or probable. Allowing these as steps in arguments is controversial because normally we think that diagrams involve non-propositional data which is related to representational accuracy and not to truth-functionality. What proposition is the Mona Lisa? Is it true or false? In other words, we think pictures are representations, not propositions, and we think they are more or less accurate as representations, and not true or false. This is the first indication that what Norton means by “argument” is something more encompassing than what is typically meant by empiricists, who usually conceive of arguments as mere manipulations of propositions.

Now, notice the second part of the above quotation. The picture proof is an argument that is good enough to convince “all but the sternest skeptics.” I wonder: is the proof convincing because it is a good informal argument, or is it a good informal argument because it is convincing? If it contains pictorial steps, it cannot be reconstructed propositionally. This makes me think that it is a good argument because it is convincing. In this case, Norton agrees with the Wittgensteinian idea that the important thing about proofs, which are arguments, is that they are convincing to the community. If “convincingness” plays a role in the identity of arguments, this would be another substantial and overlooked fact about the argument view.

Another detail concerns the role played by the content of a thought experiment. Does the content matter for identifying a thought experiment, or is it only the form? Well, certainly not all of the content matters. According to Norton, thought experiments necessarily “invoke particulars irrelevant to the generality of the conclusion” (1991, 129). In this way, they resemble real experiments which are inherently messy. This distinguishes them from standard logical arguments, which do not add superfluous detail. This is why for Norton, the conclusion of the thought experiment does not rely on the irrelevant particulars invoked in the premises (1991,
But what about the relevant particulars? These surely include more than the mere form of the thought experimental inference, because two thought experiments which can equally be reconstructed as instantiations of modus ponens are only equally valid—they are not equal. This is especially true for inductive arguments, whose cogency depends on the content of the premises. There is an important difference between the inference from a few black swans to “all swans are black” and the inference from a few samples of copper conducting electricity to “copper conducts electricity.” Both are enumerative inductions, but the first is weak while the second is strong, even though the form of the inferences is the same; namely, from the properties of some instances to the properties of all. And Norton agrees that the content of an argument’s premises does make a difference to its identity, and thus to which argument it is, at least with respect to inductive arguments (Norton 2014). To assume otherwise leaves the argument view open to Marco Buzzoni’s criticism (2008, 66-67), that by excluding any reference to concrete particulars Norton divorces thought experiments from the world completely. If “thought experiments, stripped of any reference to concrete experimental situations, are confined to a domain of purely theoretical statements and demonstrative connections,” they will inevitably fail to serve their purpose of telling us something about the world (Buzzoni 2008, 67). The reply I am making on Norton’s behalf is that if thought experiments are arguments, some of the content of their premises does matter for their identification and individuation. Any given particular invoked by a deductive thought experiment might be irrelevant for the justification of the conclusion, and that is because what justifies the conclusion is something about the logical form of the argument. But this is not the same as saying that no particulars are necessary, or that the particulars play no role in the argument at all. In fact Norton argues elsewhere for an account according to which “inductive inferences will be seen as deriving their license from facts. These facts are the material of the inductions” (2003, 648). His account of induction does not separate factual and formal aspects of inductive inferences, so the material content will be important indeed for individuating arguments (2003, 647).

In sum, Norton identifies thought experiments with deductive, inductive, abductive, diagrammatic, and informal arguments. The identity conditions for these arguments are exactly the identity conditions for thought experiments, with the added condition of equal content. A type/token distinction (inspired by Bishop 1999) will be helpful here: every thought experiment is a type of argument. The type is determined by its logical form in the case of deductive thought
experiments, and the relations between the facts responsible for cogency in the case of inductive thought experiments. But the specific (token) argument that it is depends on the form as well as the rhetorical dressing of the thought experiment (order of premises, etc.).

3 Identifying Arguments

Our ability to identify the argument instantiated by a thought experiment in every case is supposed to be a surprising fact that calls for explanation. If we cannot identify the corresponding argument for some thought experiments, then this fact disappears and we lose one of the two main motivations for the argument view. This is recognized by Norton, according to whom new information gained by a thought experiment “is elicited from information we already have by an identifiable argument, although that argument might not be laid out in detail in the statement of the thought experiment” (1991, 129, my emphasis). This section explores in what sense Norton can claim that the arguments instantiated by thought experiments are identifiable.

The reconstruction thesis, first presented in Norton (1996, 339), suggests how to reconstruct thought experiments as arguments. Norton claims that “all thought experiments can be reconstructed as arguments based on tacit or explicit assumptions.” The idea is that for every thought experiment, we can reconstruct it as an argument by locating the premises and conclusion. While the conclusion is usually explicit, we are not always so lucky with the premises, which may be found within the thought experiment itself, in some background theory, common sense, or a logical consequence of one of these. Presumably, if we can identify the premises, the logical relations between the premises and the conclusion will be determinable, and the argument will have been identified. We can see that this is method for identifying arguments in thought experiments by looking at the actual reconstructions Norton performs. There are many good examples in Norton (1991) and (1996). In each case, Norton disentangles the writing of a scientist into standard premise-and-conclusion form by abstracting the logically relevant propositions and then considering the validity or cogency of the resulting argument.

However it might not always be so easy; Norton admits that there may be cases where a thought experiment “could not be reconstructed explicitly as an argument, because the thought experiment invokes some acceptable, inductive move, to which we only assent because of the suggestiveness of the thought experiment. The deficiency here lies in our lack of understanding
of the relevant inductive move” (1991, 142-3, note 2). This leads to an additional caveat for the reconstruction thesis: both the premises and the inferences of a thought experiment may be tacit. If some thought experiments are reducible to arguments only implicitly, we may be able to identify something as an argument without being able to state exactly which argument it is, since there are cases for which we cannot identify all the steps. We know that there are steps, but we may not know what they are, or how they got us to where we ended up. This suggests a possible contradiction. On the one hand, Norton is very clearly committed to the identifiability of some argument underlying each given thought experiment. However, we have just seen that Norton also allows that in some cases it is possible that no such argument will be identifiable since if both premises and inference are implicit, we will not be able to proceed. What does this mean for the reconstruction thesis?

There are two ways to dissolve this problem. First, we can understand the claim that all thought experiments are reconstructible as arguments as equivalent to the claim that someone possessing all the facts (including facts about logic) could identify the argument behind any given thought experiment. In other words, all thought experiments are reconstructible as arguments in principle, although not necessarily in practice. This is not satisfactory, however, because the reconstruction thesis purports to explain a surprising fact about thought experiments that is available to us now, not only to ideal reasoners with all the facts. In other words, without the fact that all thought experiments can now be identified with a unique argument, we lose the motivation to identify thought experiments with arguments that was supposed to result from the reconstruction thesis.

The other way to dissolve the problem is to see the passage above as Norton merely suggesting an unactualized possibility. There could be thought experiments with inferences that are not identifiable, but in fact there are none. Following the passage Norton states, “I know of no examples of thought experiments of this type.” With this, we apparently recover the curious fact that all thought experiments we know of can be reconstructed as arguments.

However allowing that an acceptable inference can be identified simply “because of the suggestiveness of the thought experiment format,” we lose the motivation to explain the curious fact that all thought experiments can be reconstructed as arguments. The rhetorical or even
emotional force of a thought experiment can be all the evidence we need to know that we are faced with an acceptable inductive inference. Identifiability in this sense is not sufficient for the argument view. First, many fallacies of reasoning are suggestive, and we most certainly should not take such suggestiveness to be evidence of an “acceptable” inductive inference. Second, and more importantly, this means that for Norton, an argument is no less an argument if it relies necessarily on non-logical elements for its identification. He seems to be committed to the idea that anything cognitively suggestive is an argument. What was earlier a surprising fact has been transformed into an analytic truth: since thought experiments are suggestive, and anything suggestive is an argument, thought experiments are arguments.

To summarize: in this section we asked whether we should see our ability to reconstruct all thought experiments as arguments as a fact necessitating explanation by asking whether we could indeed always carry out such reconstruction. I concluded that by allowing that some thought experiments are only reconstructible implicitly, and are only recognizable as arguments in the first place because of their suggestiveness, we lose our grip on the “fact” that we can always reconstruct thought experiments as arguments.

Having cast doubt on our practical ability to identify the instantiated argument for all thought experiments, I want to turn to Norton’s reasons for thinking that we are arguing even when we think we are performing a thought experiment. This is important, because if Norton can establish that all performances of thought experiments are necessarily arguments, then the likelihood that we will one day run into a thought experiment that cannot in practice be reconstructed as an argument vanishes completely.

4 Performing and Justifying Thought Experiments

Norton claims that “the actual conduct of a thought experiment consists of the execution of an argument, although this may not be obvious since the argument may appear only in abbreviated form and with suppressed premises” (1996, 354). He realizes that many will be reluctant to accept about this extension of the argument view. “Perhaps logic and argument ought to reign in the pristine context of justification. But in the murky context of discovery, who can say what mysterious processes lead us to our conclusions?” (1996, 356). We can, he thinks, for two reasons.
First, thought experiments “come to us as words on paper” (2004b, 51); they often make explicit assumptions, and assumptions are things you find in arguments (1996, 355). Therefore, he asks, “Is it so implausible to fill in the gaps with tacit assumptions and argumentation?” (1996, 355). This argument suggests that since some thought experiments are presented via the same medium as arguments (words), and they both rely on assumptions, the two are cognitively or psychologically equal. This is a fallacious argument; consider laboratory experiments, which can be described to us in words and rely on assumptions, but are not arguments. Or consider the imperative “Go to bed!” This is not an argument (except in the vulgar sense of the term) despite relying on assumptions and appearing in words.

Here is Norton’s second argument: “For years, introductory informal logic classes have been taught on the presumption that much of thought is actually tacit argumentation, that it can and should be made explicit and that its soundness can be tested by comparison with known valid and fallacious argument schemes” (1996, 355-6). This is elaborated in a later paper as follows, “At this level of description thought experimenting does not differ from the reading of the broader literature in persuasive writing. A long tradition in informal logic maintains that this activity is merely argumentation and that most of us have some natural facility in it” (2004b, 51).

In response to this argument one wonders: since when do we base metaphysical claims about human cognition on what introductory logic classes presume? Norton’s argument is either an appeal to tradition or to the insight of those writing introductory textbooks in informal logic, neither of which justifies his claim on their own.

To conclude, Norton’s two reasons for thinking that the performance of a thought experiment is cognitively identical to the execution of an argument are unsatisfactory. Let us therefore turn to the other main support for the argument view: the reliability thesis.

The reliability thesis (1996, 356; 2004a, 1143) also depends on a surprising fact that requires explanation: thought experiments are always exactly as justified as their corresponding arguments. Norton claims that this coincidence is best explained if thought experiments are arguments. Only this identification allows us to avoid the situation in which we have two methods that always provide perfectly equally justified outputs. Such a situation would be acceptable if not for the fact that we can only determine the level of justification for arguments,
and not for thought experiments. Thought experiments are therefore superfluous because their justification piggybacks on the arguments they disguise.

The first thing to note about this “coincidence” is that it is only striking if all thought experiments really are justified to the same extent as their reconstructed arguments, which in turn assumes that all thought experiments can be reconstructed as arguments. In other words, the reliability thesis relies on the truth of the reconstruction thesis, which we have already found reason to doubt.

However let us assume the reconstruction thesis for the sake of argument in order to consider the arguments that Norton develops for identifying thought experiments with arguments.

5 The Move to Logic

The “most striking coincidence” (1996, 356) of the always equal justification of thought experiments and their logical reconstructions is supposed to serve as evidence that thought experiments are arguments. According to a weak reading of this thesis, there are many possible explanations for this coincidence, and the identity of thought experiments as arguments is merely the best of them. According to a stronger reading, the identity of thought experiments as arguments is the only way we can make sense of the coincidence.

There is no argument for the strong reading in Norton (1996). However, in Norton (2004a, 2004b) he does seem to hold a version of the strong reading. His claim is no longer that the coincidence mentioned above suggests that thought experiments are arguments. Rather, Norton tells us that thought experiments are reliable; and “this is only possible when they are governed by some very generalized logic” (2004a, 1140, my emphasis). Norton suggests that such logics “are most likely the familiar logics of induction and deduction,” and adds that “other accounts can offer a viable epistemology only insofar as they already incorporate the notion that thought experimentation is governed by a logic, possibly of very generalized form” (2004b, 45).

He seems to be saying that a cognitive process is reliable only if governed by a logic. Norton believes that if he can show that thought experiments are governed by a logic, then they are arguments. Of course, Norton needs to show that anything governed by a logic is an argument. In
the rest of this chapter, I will question whether being governed by a “very generalized logic” is really the same thing as being an argument.

6 A “Very Generalized Logic”

For Norton, “to count as a logic in this most general sense, the specification of the admissible forms must admit some systematization; for practical reasons we would hope they are tractable and communicable” (2004b, 43). A logic is the specification of such a systematization, because logic gives the form, and not the content, of some system. Norton concludes, “As far as I can see, this systematizable distinction of form and content is all we need to say that we have a logic in the most general sense. One might be tempted to impose further restrictions. But naturally arising restrictions seem to me to be too restrictive” (2004b, 53).

He says that any attempt to confine logic to 1) a truth preserving system, 2) a formalized system, or 3) a system designed to avoid contradiction will fail because adopting any of these criteria would disallow existing logical practices (2004b, 53). Therefore, we should limit our description of generalized logic to the study of the formal properties of a set of elements, although the logic need not be formal itself. Norton is claiming that whenever there is a set of elements with formal relations between those elements, we could develop a logic that governs it. In this case the question becomes: are all manipulations of the members of a set arguments, or only some of them? Even if we suppose that there is a logic so general that it governs all possible formal manipulations, such a logic would not tell us on its own which manipulations should count as arguments, unless all do. If all manipulations count, then arguments are no more interesting than making sentences. If arguments are a special kind of manipulation, then for Norton to be correct, whatever makes arguments special must also be present in thought experiments. What might this additional feature be?

To answer this question we need to find out what makes an argument an argument, given its being governed by a very generalized logic. Then we must check to see if this property or set of properties is also found in thought experiments.
7 A Broader Conception of Argument

Starting in 2004, Norton’s conception of an argument is broader even “than the one usually invoked in logic texts” (2004b, 52). To outline such a conception, I will look at some conceptions of argument found in logic texts to see how much more general we can get. Here are three:

An argument is a set of two or more sentences, one of which is designated as the conclusion and the others as the premises. This is a very broad notion of an argument. For example, it allows us to count the following as an argument: Herbert is four years old. [Therefore] The sun will shine tomorrow. This is, of course, a paradigm of a bad argument. (Bergmann et al. 2004, 9)

An argument is a collection of statements divided in premises and conclusions…our definition allows for arguments with more than one conclusion…our definition does not require that anybody ever takes the premises to be grounds for believing the conclusion. It doesn’t require that the argument was ever used, uttered, or thought by anybody, nor that there be any relationship between the premises and the conclusion…There are various advantages to using such a broad definition of ‘argument.’ The most important is that it disentangles the questions “Is this an argument?” and “Is this argument any good?” because there is no minimum standard or plausibility a group of statements must meet to count as a real argument. (Bell et al. 2001, 1-2)

For our purposes, it is convenient to see an argument as a sequence of sentences, with the premises at the beginning and the conclusion at the end of the argument. An argument can contain a number of smaller steps, subarguments, whose conclusions serve as the premises of the main argument. But we can ignore this complication and similar complications without missing anything essential. (Gamut 1991, 1)

These conceptions are quite broad already. The only essential thing is the division between premises and conclusion(s). There are three main differences between the above notions and Norton’s. First, we saw above that perhaps Norton would deny that what is divided into premises
and conclusions are statements (or sentences), because there is a role for diagrams in arguments. In this respect, Norton’s notion is broader.

The second concerns the relation between a set of premises and their conclusion. Norton says “the efforts of logicians to codify new inferential practices ensures that the presently familiar logics of deduction and induction will in practice suffice as the logic of present, reliable thought experiments” (2004b, 52). Present-day logicians define present-day reliable cognitive practices. However, “Thought experiments are arguments, but not because thought experimenters have sought to confine themselves to the modes in the existing literature on argumentation; it is because the literature on argumentation has adapted itself to thought experiments” (2004b, 64). Logic evolves to account for successful modes of thought. If we reason in a new and successful way, logicians will develop a logic to explain when and how it works. Therefore, the notion of argument should be extended to anything that expresses a logical form that falls under any present or future logic. This seems like a broader conception of argument, but it is not, since those notions are not concerned with the inferences that connect the premises to the conclusion: there need not be any connection at all. Bell et al. even allow for multiple conclusions to be drawn from a single argument, something that Norton does not (2004b, 63-64). So while Norton allows for valid inferences that only future logicians will be able to classify as valid, those arguments will still have premises and conclusions, which is all that is required by current textbook definitions. This is an important difference between what counts as a logic for Norton and for introductory logic texts, although it is not one that broadens the concept.

The third difference concerns the way we evaluate arguments. Again, Norton appeals to “evolutionary considerations” to allow for argument forms that may be justifiable only by future logics. This seems neither to narrow nor broaden the notion of argument, since all notions admit the possibility of bad arguments. New ways of justifying arguments will not expand the set of arguments. However, it does in fact broaden the notion of argument because of the following implication.

For Norton, knowing that a thought experiment is “good” means being able to identify some mark that separates it from bad thought experiments. This mark must be internal, that is, we do not need to know who performed it, or when/where it was performed. This is because there is a
distinction between form and content, even in thought experiments. This fact is what ensures that thought experiments fall under a generalized logic (2004b, 13). Also, “The mark is just that the thought experiment either uses an argument form licensed by a logic or can be reconstructed as one” (2004b, 54). To me, this is the crux of Norton’s account. Assuming that all thought experiments respect a distinction between form and content, what we must do is see what happens when a given thought experiment is reconstructed into its formal elements. If this is done, and the thought experiment turns out to possess a form that logic has identified (or could identify) as a good one, then the thought experiment is an argument.

There are four major implications of this. First, there are two primary features of arguments: their distinction between form and content and their instantiating known and approved logical schemas (their mark) (2004b, 52). Their mark, that is, their instantiation of a logical form, is an aspect of the formal properties of the argument, and not an aspect of its content. For this reason, Norton believes a thought experiment works “independently of its history.” The mark in question is not something “about the context in which it was proposed” (2004b, 53). This is in tension, however, with what we saw earlier about convincingness being important to the individuation of an argument. What is convincing changes with context. Norton must therefore give up one of two claims: contextual convincingness as an individuation criterion for arguments, or the independence from context of an argument’s logical mark. If he gives up the first, future logicians might not be able to recognize which inferences are at heart deserving of formalization and codification. However he cannot give up the second since it describes a core feature of thought experiments and arguments. It seems clear that he should give up the first. However, I think instead he should give up the idea that the mark will be completely internal to the thought experiment. As we saw above, the content of the thought experiments matters for which argument it is, and the content of the terms or representations in the premises and conclusion will depend on the community who performs the thought experiment. The mark therefore, will not merely be a property of the logical form of the argument, but a property of both the form and the material of the inference.

Second, if a (good) argument is no more than something that obeys the distinction between form and content and instantiates a form approved by logic, a (good) argument or thought experiment should be identifiable and perhaps even evaluable in advance of its being performed. This is an
interesting idea, but one that leads to trouble. First, without the context of its performance, we cannot say whether a thought experiment is really justified, because the meanings of the terms in the thought experiment can change. Second, if a good argument is just something with a form/content distinction that is approved by some logic, then a bad argument is only something that admits of a form/content distinction which is not approved by a logic. In this case, everything in the universe, actual, possible or impossible, is a bad argument if it is not a good one, since everything has a form/content distinction. This makes the circularity of the reconstruction thesis even more obvious. Also, since a logic is just some systematizable distinction between form and content, then we can invent two logics, one that makes some inference a good argument and another that makes the same inference a bad argument. To choose between competing logics, we will need another mark that distinguishes good logical systems from bad ones. If such a mark is a property of the form of the logic, then we are in for an infinite regress. If it is not a formal property that marks a good logic, then why must it be a formal property that marks a good thought experiment (or argument)? This idea is taken up in the next chapter.

Third, for Norton, being justified by a logic is the same as being an argument because something cannot be justified by a logic without being an instance of one of the forms licensed by that logic. (I am translating “If thought experiments can be used reliably epistemically, then they must be arguments” into Norton’s more considered language following this statement in his 2004b). This is another sense in which the notion of argument can be said to have been broadened: if something is cognitively successful, it is an argument. To disprove this claim we need something that is cognitively successful but which is not an argument. But this is impossible, given that success for Norton means instantiating a form licensed by logic. This makes the reliability thesis circular.

Finally, inductive arguments, for Norton, admit no distinction between form and content. Norton has recently put forward a theory of inductive inference he calls the “material theory of induction” (Norton 2003, 2005, 2010, 2011, 2014, forthcoming). “If one adopts a material theory of induction, one no longer separates factual content from the rules of inductive inference. The problem of justifying some particular induction is replaced by the straightforward task of justifying the facts that warrant it” (2014, 672). If we stop separating the rules of inference from
the “factual content” then we have eliminated the distinction between the form and the content of inductive arguments. This means that according to Norton’s own definition of an argument, inductive arguments are not arguments, and inductive thought experiments are not either. A closer look will make the point clearer.

The material theory of induction is to be contrasted to the formal theory of induction, which seeks universal rules that justify types of inductive inference. Norton’s material theory denies the possibility of these universal rules. Instead, each induction must be evaluated separately based on the facts upon which the inference relies. Opponents of Norton (Sober 1988, Okasha 2005, Worrall 2010, and Kelly 2010) claim that the material theory can be reduced to the formal theory. Norton replies that this claim “depends on a tacit and unwarranted assumption: that the relations of inductive support among facts are organized in a strictly hierarchical structure” (2014, 672). He rejects this assumption, arguing that inductive facts are not organized in a hierarchical structure but that “the warranting relations cross over in many directions” (672), interwoven even with provisionally adopted facts (673). Science is just such an interwoven mesh of facts that support each other inductively, according to Norton, who again, does not allow for any distinction between form and content in successful inductive inferences. Therefore neither inductive arguments nor inductive thought experiments can be identified by their distinction between form and content, or compared in terms of justification by their instantiating successful logical schemas, because there are no general logical schemata for inductive inferences.

Also, since there are different webs of facts that make up science at different times, we cannot have the “mark” of a good argument (thought experiment) being completely independent of its history when that argument (thought experiment) is inductive.

Let me emphasize this point one more time. According to the formal theory of induction, “valid inductive inferences are distinguished by their conformity to universal templates. They may be simple, such as the template that licenses an inference from some past A’s being B to the conclusion that all A’s are B. Or they may be more complicated, such as the requirement that degrees of inductive support conform to the probability calculus” (Norton 2014, 673). Norton denies the feasibility of this project. And this denial contradicts the notion that inductive
arguments are individuated or evaluated based on their having a form/content distinction or fitting into “an argument form licensed by a logic” (Norton 2004b, 54).

8 Conclusion

Given the notion of argument that I have extracted in this chapter, it seems impossible to prove that something is not an argument in Norton’s sense (unless we could show that the completed future logic would be incapable of characterizing that something). This is what is behind sentiments like this one expressed by Brown: “Norton says that thought experiments are often disguised, not explicit arguments. So the real claim is that they can be reconstructed along his empiricist lines. Existence claims like this are devilishly difficult to defeat. I doubt that an actual refutation could ever be delivered” (1992, 275). We saw in this section just how impossible such a refutation would be, as long as we accept the circular notion of an argument as anything justifiable by a generalized logic.

However we also saw that Norton’s identity conditions for thought experiments and arguments conflict with his newest work on induction. Assuming Norton will not give up his material theory of induction, he must then give up his work on the mark of a thought experiment being a logical feature that is independent of context. However this will be difficult, as it is a fundamental feature of his account.

In the next chapter, I want to consider another serious problem for Norton’s account, which is the notion of being governed by a very generalized logic. I concluded in this chapter that such a notion would not on its own suffice to produce an individuation criterion for arguments (or thought experiments), or a way to identify them. What I will discuss next is whether an empiricist like Norton can coherently claim that logic is the kind of thing that governs at all. If I can cast doubt on this underlying assumption of Norton’s, I will have shown that neither the argument for his view (Chapter 2) nor one of his foundational assumptions (Chapter 3) are convincing as they stand.
Chapter 3
Thought Experiments and Logic

There is no official book of rules for the direction of the scientific mind.

(Burgess and Rosen 1997, 208)

According to Norton, it does not matter whether we imagine Newton’s bucket made of wood or glass or lunch meat, or whether we first encounter this thought experiment in 1730 or 1930. The defining “mark” that tells us whether it is justified is completely internal; it is a structural feature shared by other similarly successful thought experiments (2004b, 54).

We saw in Chapter 2 that logic, for Norton, is what allows us to distinguish between justified and unjustified uses of an inferential practice. For instance, he claims, “If we are to recognize the logic as delimiting the successful thought experiments, there must be something in the logic that evidently confers the power of a thought experiment to justify its conclusion” (2004b, 54, my emphasis). Logic allows us to delimit justified inferential practices because logic confers justification on them.

In addition to claiming that logic “governs” inferences (2004b, 45), Norton also claims that it actually guides us when we make justified inferences: “we should expect the schemas of this logic [that governs thought experiments] not to be very complicated, so that they can be used tacitly by those who have the knack of using thought experiments reliably” (2004b, 54). Norton is not alone in this. Immanuel Kant believed that transcendental logic provided the rules by which we must think, if we are to think at all. Gottlob Frege concurred, calling the laws of logic “laws of rational thought.” Gilbert Harman disagreed, arguing that logic cannot provide laws of thought since logic changes over time, while laws of thought presumably do not (1986). Hartry Field on the other hand has tried to re-establish the connection between logic and rationality as recently as 2009. I do not have the space to address the literature on logic as being normative for rational thought. Instead it is my goal to refute the specific version of this thesis present in Norton’s account.
I will consider the two main ways an argument is usually justified: deductively and inductively. I begin with deductive arguments.

When we ask what makes a deductive argument or inference justified, we do often answer by reference to logical structure. For instance, an argument might be justified because it is an example of modus ponens. I will argue that such classifications are only a form of justification by proxy. What we mean when we say that an inference is justified on the basis of its logical form is that it is justified because it is relevantly similar to other successful inferences. But what makes these inferences successful to begin with cannot therefore be something logical. To show this is the goal of the present chapter. Instead of being committed to the claim that logic can confer justification to inferences, I will argue that for an empiricist like Norton, logic should be understood as a language useful for representing objects. Once this is allowed, Norton loses the support he claimed from the reliability thesis.

I begin by proposing an empiricist-friendly account of the applicability of logic. This account is analogous to the “mapping account” in the philosophy of mathematics (see Baker 2003; Balaguer 1998, 144; Brown 2008, 2012; Colyvan and Bueno 2011; Krantz et al. 1971; Leng 2002 and Pincock 2004a, 2004b). I will suggest that just as mathematics models but does not explain events or relations between objects/events in the physical world, logic models but does not explain events or relations between objects/events in the world of thought. Then, I will use recent work on a new type of logic meant to capture diagrammatic reasoning as a case study. I have chosen this case study because it is an example of the careful and exciting interface between logicians on the one hand, and a set of previously unschematized inferences that are antecedently thought to be justified on the other.

Before I get to the mapping account of the applicability of mathematics I should explain why I think it would be the preferred account of the applicability of mathematics for an empiricist like Norton. Empiricists explain (away) mathematical objects in several different ways. Here are three: 1) they are conceived as general terms or relations between terms, built up from manipulations with actual objects (as in J. S. Mill 1882), 2) they are admitted as objects only because they are indispensable for the operation of science (as in Quine 1951, 1976, 1980, 1981), or 3) they are considered nominally as tautological relations between definitions that we find
useful, but could dispense with in theory (as in Field 1980). Norton cannot accept the mind-independent existence of mathematical entities because entities like numbers and sets are abstract, non-causal, and therefore cannot be known by the mind unless by Platonic intuition, which Norton derides as epistemic magic (Norton 2004b, 44). This leaves Norton with several options: mathematical truths are contingent truths generated from experience, or they are useful tautologies which can be used to model events in the natural world. In both cases, something must be said about the connection between mathematics and the world that it helps us to understand.

If mathematical statements are truths generated or abstracted from sense experience (as in Mill), then we should not expect to find those statements ever serving as the ultimate justification for empirical claims. Rather, the statements of mathematics should capture and represent in formal ways truths that already obtain, in much the same way that the statements of physics depend for their truth values on experience with physical objects and events. On the other hand, mathematics might be an account of the relations between stipulations, which we would only deign explore if we found such relations useful. And how might such relations be useful? What might they be useful for? David Bostock argues that pure mathematics is “a convenient fiction that is very helpful for purposes of calculation, and helpful too as providing a vocabulary with which to express our scientific theories” (Bostock 2009, 226). If we go this road, we have a system of statements that is useful for providing a vocabulary for scientific theories, a vocabulary that helps us calculate and otherwise deal with the data received by observation and experiment.

In both cases, mathematics turns out to be a language that represents. In the first case, it is a language created to help us understand and manage experience. It passes beyond the simple representation of arithmetical and geometrical relations useful for everyday life, and becomes beautiful on its own. But whatever power it wields comes only from the initial contact it makes with experience.

According to the second empirical interpretation, under which fly the banners of eliminativism, reductionism, nominalism and fictionalism, we characterize mathematics again as a language that captures or represents entities with which we are concerned. Here too we need
representation relations that specify which parts of our mathematical language can correspond to which parts of experience. The mapping account is consistent with both interpretations, and therefore with all the main empiricist accounts of mathematics that might appeal to Norton. It also happens to be the most popular account of mathematics by far. Since Norton claims we should take empiricism as default because it is the most popular position in the philosophy of science (2004b, 50), perhaps he will allow me to take the mapping account of mathematics as default for the same reason.

1 The Mapping Account of the Applicability of Mathematics

Colyvan and Bueno characterize the mapping account as follows: “According to the mapping account of mathematical applications, the explanation of the utility of mathematics is no different from explaining the utility of street maps. The idea is that there is some structure-preserving mapping between the world and the mathematical structure in question” (2011, 346-347). In other words, we represent physical relations, properties or entities using mathematical relations, properties or entities, and then we reason analogically about the former by manipulating the latter. Of course, the mapping between the mathematical structures and the physical structures may be quite complicated. The relation might be one of isomorphism, homomorphism, monomorphism, or epimorphism. But there is also the problem that some parts of a mathematical structure will usually be disanalogous to its target physical system (like the negative roots of a quadratic equation that represent an instance of projectile motion). So far, we are still in the dark when it comes to explaining in general how we know which aspects of a mathematical system represent, and which do not. Colyvan and Bueno attempt to solve this difficulty in the following way.

They argue that the main reason we apply mathematics is to make new inferences possible. By analogical reasoning through mathematics, we are given the tools to make many kinds of inferences that we could not have otherwise made. To capture our ability to make new inferences about the world using mathematics and explain how we know which parts of a mathematical structure refer and which do not, they propose a system similar to R.I.G. Hughes’s Denotation, Demonstration, Interpretation account of the way we reason about physical systems using
models in science (see Hughes 1997). Their account is the Immersion, Derivation, Interpretation account. The key is the third step: we interpret (or re-interpret) the mathematical analogy after making inferences about a target system. This creates a semantics for our mathematical model.

According to the mapping account, mathematics is a system that does not cause, explain or justify the physical structures which it models. Rather, it represents them. Mathematics itself causes or explains physical events no more than Hamlet causes or explains the events of a tortured young man whose uncle has just murdered his father and married his mother. This is the case even if we use Hamlet to make true predictions concerning the man, who goes on to see ghosts, stage a play, and experience an existential crisis (Brown 2012, 10).

There is much to be said for this view. For one, Brown notes that it explains how it is that mathematics seems to be at the heart of some scientific revolutions, while empirical science has never returned the favour (2012, 13). Brown provides the following example. While the discovery of non-Euclidean geometries opened up new modelling possibilities for physicists like Einstein to develop General Relativity and thereby counter Newtonian physics, the discovery of non-Euclidean geometry itself did not, and could not, have affected Newtonian physics. Likewise, Relativity Theory (or any other physical theory) cannot affect the validity of Euclidean geometry.

This mapping account of mathematics should be attractive for Norton. To repeat, thought experiments are identified with arguments that are deductive or inductive (which includes informal, abductive, etc.). Deductive thought experiments merely rearrange data. Since mathematics is a deductive system, it too must merely rearrange data. The reason we bother with mathematics at all is because those rearrangements are useful or necessary to present or represent important features inherent in the data. By doing this, we can reason about the data more efficiently.

On the other hand, thought experiments can also be inductive arguments. And as we saw above, Norton’s material theory of induction says that logical schema are not what justify inductive inferences in science. What justifies these inferences are rather the things that figure into those schema. So Norton should subscribe to an account that makes sense of the privileged position of
logic and mathematics in science, without according it a substantive justificatory role. This account does that.

2. The Mapping Account Applied to Logic

Just as physical objects and their relations can be modelled using mathematics, mental objects like propositions, ideas, concepts, and their relations can be modelled using logic. Of course other sorts of objects can be modelled by mathematics or logic as well; nonetheless I want to focus on physical objects and mental objects because they help bring out the analogy between mathematics and logic. Just as the metaphysical status of physical objects cannot form an obstacle to their being reasoned about using mathematics, it does not matter whether we characterize mental objects as ontologically abstract or patterns of neural activity (or something else) to reason about them using logic.

Domains, possible worlds and proposition variables are mental objects whenever we think about them. But in logic, they are not mental objects but logical terms with definitions and roles in formal theories, and they are causally (though not explanatorily) inert. The logical term POSSIBLE WORLD is different from the mental thought “possible world” (expressed in an idea or as a concept). In this context, Norton’s claim that the only criterion necessary for the existence of a logic is the distinction between form and content is relevant. At least in the case of deductive logic, Norton must uphold that there is one type of thing (the form) that does something (represents, governs, categorizes, etc.) another type of thing (the content) which fits into the formal schemas. Items like possible worlds can be part of the form or the content, but in the rest of this chapter I want to discuss them in their formal sense.

The mapping account of logic proposes that logical systems make representations of mental objects. These representations are manipulated according to established rules, and the output can be interpreted back into some claim about mental objects by reversing the semantic conventions that transformed the mental objects into representations in the first place. Logical manipulations allow us to draw conclusions in this way about the relations between mental objects (like propositions and concepts) by how their representations are related in a formal system. These conclusions are logically justified insofar as the representations are good ones, that is, insofar as
the logical terms capture the salient aspects of the target mental objects, and the logical relations capture the salient aspects of the relations between the mental objects.

Here are a pair of examples that illuminate the relation between the mapping accounts of mathematics and logic. The first is from Brown (2008, 52). We imagine a number of bodies with weight that we wish to represent using mathematical items, say, real numbers. Consider the relation *physically weighs the same or less than*. This relation can be defined operationally using an analog or digital scale. There is also the relation of *physical addition*, as when we combine two balls of putty. There are mathematical relations that we can use to model these relations, namely, “≤” and “+.” We set up a representational relation between the systems using rules such as the following:

RULE: If one object *physically* weighs the same as or less than another, then the real number associated with the first object is *mathematically* less than or equal to the number associated with the second.

RULE: the result of the *physical* addition of one object to another is equated with the real number associated with one object *mathematically* added to the real number associated with the other.

Brown argues that it would be folly to think “≤” or “+” were in anyway *responsible* for the output of physical weight measurements. It is assumed that similar representation relations could be created for each mathematical relation, although it gets more complicated. For instance, if instead of modelling weights we were modelling velocities, we would have the option of Newtonian or relativistic addition. But whatever model we choose, the mathematical relations do not cause, license, or govern the objects and events of the target systems.

Now here is a logical analogue. We wish to represent a set of mental objects and their relations using the items of first order propositional logic. First we represent beliefs (mental propositions to which we assent) as sentences or sentence variables. Then we define *psychological implication* as the psychological impetus we experience when one proposition causes us to believe another. We represent this logically using the symbol “→,” defined according to its truth table definition. Then we define *psychological negation* as belief in the denial of a proposition.
We may represent this logically using “~,” again defined classically. We then set up a representational relation using rules such as:

RULE: If one belief psychologically implies another, then the sentence variable chosen to represent the first belief logically implies (⇒) the sentence variable that represents the second.

RULE: The result of psychologically negating a belief is mapped to the result of applying the operator ‘~’ to the sentence variable used to represent that belief.

And we may use this simple logical system to represent actual beliefs and many of the relations between them. And just as in the case of mathematics, “⇒” and “~” do not cause, license or govern mental beliefs or inferences. Again, we could represent the same inferences using different systems, for example with higher order, modal, or multi-valued logic, or we may prefer strict implication in place of material implication. But this would not change the nature of the relationship between the logical and target system.

The mapping account applied to logic provides the same affordances it does in the case of mathematics. First, it explains why developments in logic or set theory may provide new modelling possibilities for other fields, while those same fields do not produce revolutions in logic. The doctoral dissertation of Emil Post shows that there are no revolutions in logic, since any non-classical logic will always preserve all classically valid inferences (Post 1921; and see Jennings and Scotch 2009, 17). Second, this understanding of the applicability of logic is either consistent with definitions of logic made by logicians, or is presupposed by them. Barwise and Hammer define a logical system as “a mathematical model of some pretheoretic notion of consequence and an existing (or possible) inferential practice that honors it” (1996, 51). Notice the word “model.” L.T.F. Gamut explains the importance of logic as two-fold. First, “logic contributes systems which give a precise description of a number of expressions which, because of their importance to reasoning, cannot be ignored in a linguistic theory of meaning” (1991, 9). And second, “logic contributes methods and concepts useful for the analysis of expressions and constructions” (ibid.). Bergmann et al. say in their widely used textbook that logic should be studied because through it we “learn to symbolize natural language sentences...in a formal language” (2004, 6). While Gamut’s “describing” and “analyzing” and Bergmann et al.’s
“symbolizing” are not as strong as “modelling,” I hope my point is apparent: the task of the logician is to capture (describe, model, represent) something that is already out there. And this notion of one thing capturing another is also present in Norton’s description of a very generalized logic as well, which captures successful reasoning practices using logical schemas.

While it is nice that the mapping account is consistent with what logicians say about logic, there is another consilience that impresses me more. This is with Lewis Carroll’s article, “What the Tortoise Said to Achilles.” Carroll shows that we need to be competent with inferences like modus ponens before we can formulate them logically, understand their formulizations, and prove or disprove them. Competence with modus ponens is independent of and prior to any logical formulation or proof of modus ponens. Peter Winch writes, “the actual process of drawing an inference, which is after all at the heart of logic, is something which cannot be represented as a logical formula … Learning to infer is not just a matter of being taught about explicit logical relations between propositions; it is learning to do something” (1958, 57). This insight accords with the mapping account of logic, which portrays logic as representing our successful practices. It does not constitute, determine or govern them.

One objection to my extension of the mapping account to logic might concern an apparent disanalogy between logic and mathematics. While mathematical structures may be used to model physical systems, they do not tell you whether those systems are good or bad. Logic, on the other hand, seems inherently normative. It is in the business of telling us which inferences are good or bad. To reply, consider logic and mathematics as applied versus pure. Applied mathematics certainly does help us label good and bad uses of a mathematically represented system, where “good” and “bad” are understood with respect to specific aims (like sending a satellite to Saturn). In applied logic, it is the same. The difference is that in applied logic, our aim is usually truth-preservation, whereas the aims of applied mathematics are legion. So while it might appear that applied logic is more normative than applied mathematics, this is only because it is applied to achieve a single goal, where mathematics achieves many.

Turning to pure systems, we see that pure logic is just as normative as pure mathematics. Pure logic focuses on proof and consistency, not for everyday reasoning goals, but for the same reasons that mathematicians use and shape those same concepts. Proof and consistency are
interesting notions in their own right, as are questions concerning which theorems are deducible from which systems and which axioms capture the greatest number of intuitively valid inferences. In both cases, norms of right logical or mathematical conduct are established and enforced, while in neither case are representations tied normatively to the world.

Since Brown is an adherent of this view, I should explain how it is possible for both Brown and Norton to employ it. The short answer is that mapping accounts are only Platonic if they map Platonic entities, which they need not. Brown notes that we can understand mathematics as a representation of physical objects or as a representation of their (abstract) properties (2008, 54). When we represent the physical addition of the weight of two objects as $7 + 3 = 10$, there are at least two ways to interpret this. We might be saying that one object of weight 7 added to another object of weight 3 yields an object of weight 10. Or we might be saying that an object with the property “having weight 7” added to another with the property “having weight 3” will yield an object with the property “having weight 10.” Therefore the mathematical system could be a model of the addition of physical objects, or of abstract properties. These two options correspond to empiricist and realist accounts of the laws of nature respectively. Empiricists say that a law of nature is merely a statement that keeps track of the regularities between actual objects, while realists about laws of nature claim that laws are statements that express relations among properties. Brown will stick to the latter interpretation, and Norton the former.

This distinction holds equally well when applied to logic. Either the “laws of logic” apply to mental objects themselves (which need not be abstract), or they apply to the abstract properties of those objects. An empiricist will say that laws of logic are only representations of regularities concerning mental objects and the relations between them. An empiricist might interpret the logical axiom $\sim P \leftrightarrow P$ in the following way: “negating a negated proposition will yield the original proposition and vice-versa.” The realist about laws of logic will interpret the same axiom this way: “Applying the property ‘negate’ to a proposition already possessing that property will yield a proposition that does not have the property ‘negated,’ and vice-versa.” I make this point simply to show that this view of the applicability of logic can be interpreted in a way that is consistent with both Brown’s Platonism and Norton’s empiricism. If this were not the case, I would be begging the question.
To summarize this section, I have gestured towards an account of the application of logic that I think is an attractive one for empiricists like Norton. In the next section I will begin to argue that according to this account, it makes little sense to say that logic is what guides or justifies good inference making. I will illustrate this claim using a case study before making some more general remarks.

3 A Case Study: Operational Constraints

Many new developments in logic begin with attempts to capture and model the different ways we form good inferences. Examples include probabilistic and paraconsistent logic. Once we have identified a class of good inferences, we may try to represent them formally. But this act of representation does not confer any justification that was not already there. To see this, I would like to consider a less familiar but still recent example of cutting-edge work in logic: the attempt to capture the valid inferences that we make when reasoning with diagrams. One reason I choose this particular case study is that Norton must show that diagrammatic reasoning (such as that employed in Brown’s picture proofs) is merely argumentative. As we saw in Chapter 2, the way Norton would like to do this is by showing that such reasoning is governed by logic. As it happens, such a project is already well underway (see Barwise and Allwein 1996), and it is from this project that I draw my example.

Shimojima (1996) tries to formalize successful diagrammatic reasoning. He begins by noting that diagrammatic reasoning consists of two kinds of activities: physical operations on diagram sites (for example, drawing lines on a piece of paper), and extracting information from those diagrams (interpretation using semantic rules). He focuses on the first kind of activity, as it is here that we will find the constraints that drive the conclusions of picture proofs and other diagram-based styles of inference. There are three kinds of constraint, which Mary Hesse (1966) labels positive, neutral, and negative analogies. These obtain between the objects represented and their representations. If we were to draw squares on the sides of a right angle triangle in order to prove the Pythagorean Theorem, the colour of ink should not be something upon which we base any inferences. It is a negative analogy because it detracts from our ability to focus on what is relevant. On the other hand, positive and neutral analogies allow us to see (often very quickly)
what a formal derivation might never illuminate, for example, structural similarities. These are called “free rides” by Shimojima.

A free ride is provided from a set of information (the premises) to a conclusion by means of a specific diagram, when the following conditions are satisfied:

1. There are a number of valid operations defined by a diagram type, a sequence of which is applied to a diagram site.

2. These operations jointly realize a new fact on the diagram site.

3. None of the operations logically entails this new fact, although taken together the new fact is guaranteed because of the type of diagram being used.

4. Given the semantic convention in use, the new fact implies the desired conclusion.

The idea is that you obtain the new fact “without taking any step specifically designed for it,” and this fact leads to the conclusion (Shimojima 1996, 32; italics original).

Here is an example. We are asked to check the validity of the following syllogism:

(i) All Cs are Bs.

(ii) No Bs are As.

(iii) (Therefore) no Cs are As.

From (i) and (ii) we draw the following Euler circles:

![Euler Circles](image)

Fig. 5: Euler Circles to Reason through a Claim
None of the individual operations involved (drawing and labelling the three circles) entails the realization of the following fact: the A-circle and the C-circle do not overlap. Yet, together, barring any strange topological anomalies of the paper or other interferences which are logically possible but not captured by the standard use of Euler circles, these operations force that fact. And given that fact, and the semantic conventions that go into the use of Euler circles, we conclude that (iii) is correct: no Cs are As. Given the diagrammatic manipulations, we were able to derive a conclusion that we could represent using logic. To show that these inferences can be captured using logic, I would have to get into the technical details of Shimojima’s treatment of free rides, which do not need to be summarized here because even if Shimojima’s effort is unsuccessful, two things are clear. First, there is a definite distinction between that which does the inferential work, and the logical structures developed to represent it. Second, Shimojima’s aim is to represent usefully the former with the latter.

On the one hand, we have operation sites (for example, a piece of paper) and the methods and effects of operations—or “actions”—performed on those sites (for example, the rules for that kind of diagram, the drawing of lines on the paper, etc.). These are the objects to be modelled. On the other hand, we have the sets of variables used to represent diagram sites and actions, and a host of binary and tertiary relations that model the transformation of one stage of a diagram site into another via some action. This is typical of what logicians do whenever a new kind of inference is initially formalized. It is very clear that what is sought is a good representation; a system of symbols and relations that can be used to capture and then reason about objects or properties. This understanding of the application of logic conflicts with Norton’s characterization of logic as something that confers justification or governs inferences.

4 Implications for the Argument View

Norton claims that all thought experiments are precisely as reliable as the formalized schema into which we can translate them, and this is hard to accept if thought experiments are not arguments (Norton 2004b, 52). However, the reason a thought experiment and its logical formalization are equally justified is because logicians have done well to create a representation system that captures successful inferences. Logic is good at reflecting back to us what we are doing, but it does not confer any additional justification that was not already there.
For this reason, Norton’s appeal to logic as the justification of thought experiments (or arguments) is not successful. Logic may sharpen our intuitions, and it contains a great wealth of knowledge that we can refer to if we want to see whether something fits a certain schema that we know to be a good one, but logic gives us (at best) only a picture of something else. Ultimately, it neither governs nor justifies nor guides us in making inferences. In this case, the reliability thesis loses its pull, because it assumed that thought experiments and arguments would always be equally justified given that their justification came from the same place: their logical form. But this assumption has been shown to be untenable. Neither arguments nor thought experiments get their justification from logical form, therefore this cannot be what unites thought experiments with arguments.

Before moving on I should say something about inductive thought experiments. Norton claims that an inductive thought experiment is justified when it instantiates a form that is marked by logic as a good one. Yet as we saw, Norton also espouses a material theory of induction according to which an inductive inference is good not because of anything to do with logic or formal properties, but the things that figure into the inferences.

Given Norton’s theory of induction, in the case of inductive thought experiments, the reliability thesis is trivial. It is not a curious fact to be explained that thought experiments have the same strength as their rewritten arguments, because if they do have equal strength it is a result of the properties of the phenomena featured in the premises, not the form of the inferences or the logic we use to represent them. In other words, even if inductive thought experiments can always be rewritten as inductive arguments, and even if the two are always equally justified, this would be because the thought experiment and the argument concern the same things, not because thought experiments and arguments are identical.

So just as for deductive thought experiments, considering the nature of inductive logic and its relation to what it represents suggests that thought experiments and arguments are not equally justified (if they are) because logic justifies them both. What justifies an inductive thought experiment are material and not formal elements, and so reconstruction is only helpful insofar as it clarifies what is going on. Such reconstruction is helpful, but not the source of justification. Nothing about logic governs or guides the inference-maker. This is similar to the situation with
laboratory experiments, which also consider material objects and events and the relationships between the two. What justifies our confidence in a laboratory experiment are the features and processes instantiated in the material, brought out by intentional actions performed on material representations of some target system. What justifies our confidence in the laboratory experiment reconstructed as an inductive argument is the fact that the events of the laboratory experiment were accurately represented in the logical reconstruction.

A simple example. We ask a toddler, can a square block fit into a circular hole? The toddler might place a nearby square block into a large circular hole and thus answer our question. But the toddler might also perform that action and then provide the following report: “I put a square block into a circular hole, and if I can do that, then there are at least some square blocks that fit in circular holes. So yes, square blocks can fit into circular holes.” (We are imagining a very philosophical toddler). In the material case, what makes the toddler’s action possible is the fact that there is no law-like connection between volumes of cubes and diameters of circles. Nothing prevents us from creating cubes small enough to fit through circular holes, or circular holes large enough to admit cubic blocks. Now notice that the exact same fact is also what justifies the argument given by our philosophical toddler. If the toddler had been a little more philosophical, he or she might have represented the inference in logic. But that would not have changed the source of its justification.

The same is true of inductive thought experiments and arguments. While we can reconstruct inductive thought experiments as arguments, and while the same set of material features may indeed be what justifies each, this provides no reason to believe that the thought experiment is the reconstructed argument. In all likelihood, the argumentative reconstruction of a thought experiment will depend ontologically on the performance of the experiment.

This discussion inspires both a question and a tentative answer. If logic is not at the core of what makes a thought experiment (or an argument) good or bad, what is? Logicians codify the inferences we approve of, which are the ones that work. And it is likely that we know which inferences work by trial and error. Or in other words, by observation and experiment. Perhaps then, we should consider what makes a good experiment. Here are five criteria. 1) The system to be investigated must be well-isolated, abstracting away irrelevant features of the surrounding
environment. 2) Experimental bias must be eliminated. 3) Sources of error must be understood and accounted for. 4) Instruments should be calibrated as best we can. And 5) we should have a theory of our instruments. (These are mostly drawn from the classic Franklin 1986). These criteria may be applied to thought experiments, where we understand the instrument of observation to be the mind. For instance, 1) we can isolate irrelevant features of systems easily in our minds (this is what abstraction is). Imaginary isolation of properties and processes is easier to do than the material equivalent, and this is likely one reason thought experiments are so ubiquitous. 2) Humans are notoriously guilty of confirmation bias. But we can keep this in check by discussing thought experiments with others who do not share our expectations. 3) Inaccurate representations of a target system or misunderstanding important relations are both sources of error for thought experiments. Weak imaginations and cognitive biases are also problems which must be appreciated. 4) We calibrate our imaginations through years practicing hypothetical reasoning. In all skilled positions from electrician to tennis player, we will have considered abstract cases almost every day as part of our education. 5) Our theory of our instruments is still being developed in psychology, sociology, philosophy of mind, neuroscience, and much of the arts, so there is a great deal to work with here. This is the faintest sketch, but it should be clear how we might move forward with such an idea.

One objection is that thought experiments are not experimental because they do not intervene on extra-mental systems, which is one of the most important features of a laboratory experiment. Here are two replies to such an objection.

Intervention—while very important—is not a necessary condition of laboratory experiments. There are many examples in science where an experiment does not intervene. A classic example concerns peppered moth camouflage (Kettlewell 1955, 1956, and 1958). There are several colourings of peppered moths in England, and it appeared that the darker (and rarer) forms were becoming more common in areas downwind of factories responsible for industrial pollution that darkened the trees and killed much of the lichen cover. To see if the spread of the melanic gene was due to the inability of predators to spot the darker moths against the darkened trees, Henry Kettlewell collected moths of different colours and placed them on different coloured tree barks in a laboratory. Predators were then brought in to see which moths were preyed upon. Kettlewell found a significant statistical advantage for the darker moths when darkened tree bark was used:
they were more than twice as likely as the paler moths to avoid predation. This was careful observation of natural processes, not intervention. Nevertheless, it has all the trappings of a controlled experiment, and it established a historically important result.

Second, some homomorphic representations (for example, material models) do allow us to intervene in features of natural processes. While the aim of a scientist might be to learn about some natural system, scientists primarily do their intervening on idealized versions of these systems. We can map the strength of different representational relations on a continuum that runs from direct material reasoning (like fitting the lids on jars) through scientific field experiment, laboratory experiment, scale material modelling, computer simulation, thought experiment, all the way to reasoning with systems that use representations that have almost nothing in common with what they represent (like the dots and dashes of Morse code). When we consider the continuum of experiment, we notice that direct intervention is an ideal that typically goes unachieved. We mostly intervene on abstract and idealized versions of natural systems, not on the systems themselves. And what is a thought experimental scenario but an abstract version of a natural system? It is true that we cannot be sure that a mentally represented system will react as a natural system would, but this problem is only gradually more severe than it is for material models, computer simulations and laboratory experiments. In all these cases, we hope that our representations of natural systems preserve enough of the relevant structures of the natural systems that something new can be learned from them.

But how do we intervene on an abstract representation? A preliminary answer is given by Shimojima. There are certain actions permitted for each type of representation system, and we may use any of these on a diagram or model or imagined scenario that we like. The key to profiting from such manipulations is Shimojima’s free ride: we can perform actions on a representation, none of which are designed to produce any specific fact, and nevertheless the combination of constraints and actions performed produce a new fact that answers our question. And this method of fruitful intervention is general; it applies equally to reasoning on natural or experimental systems, or models of those systems.

This answer to the question of how we might intervene in a thought experiment needs to be filled in. To intervene experimentally is to interrupt or otherwise tamper with the time-evolution of a
law-like dynamic process. When done right, we are able to gauge how the process compares pre- and post-intervention. It is not immediately obvious how the imaginary scenarios of a thought experiment can do this. However, there are several authors who point out the dynamic aspects of mental manipulation (Miščević 1992, 220; Nersessian 2007, 143), and they emphasize the rules built into our fictional scenarios. As with the Euler circles above, there are conventions governing acceptable actions to take in different types of imagined scenarios. These will be more flexible and less obvious than with Euler circles, but perhaps looking at the way we reason with fiction might reveal some of them (Ichikawa and Jarvis 2009).

5  Conclusion

Hopefully, the factors that justify a deductive or inductive inference will show up in our logical reconstructions of those inferences. But whatever logic we use to represent good inferences will not be the ultimate source of whatever justification those inferences possess. According to the empiricist-friendly account of logic presented in this chapter, inferences are not governed by logic, and such governance is not the source of our ability to tell between good and bad inferences except by proxy. Norton claims that deductive thought experiments and arguments are governed or justified by the same logic as a way of equating them. This is mistaken, however, because thought experiments and arguments can be represented using the same logic without being identical. Indeed, thought experiments can be reliable in an experimental way, and in this case, their logical reconstructions are only justified because the thought experiments are. In other words, the reliability thesis is false.

Chapters 2 and 3 of this thesis have been directed at two of the three pillars supporting Norton’s account. The remaining pillar concerns the general empiricist principle to which Norton appeals: we should not admit magical or supernatural elements into our epistemology. If we want to uphold this principle while maintaining that thought experiments need not be arguments, we come naturally to the cognitive science of human reasoning. Cognitive science studies are often wielded by naturalist philosophers. Not all naturalists are empiricists. But all naturalists will sympathize with Norton’s stance on epistemic magic: denying supernatural elements in epistemology is precisely what makes someone a naturalist.
Chapter 4 begins with this idea. What does cognitive science tell us about thought experiments? Paul Thagard, an influential cognitive scientist and philosopher of mind, claims that cognitive science can and should be applied to thought experiments. And he argues that cognitive science shows thought experiments to be unjustified—they are epistemically dangerous and even harmful. Opposing Thagard are several philosophers who argue that cognitive science shows thought experiments to be reliable manipulations of mental models that take advantage of dependable cognitive processes. Chapter 4 uses this debate as a way of getting clear on the relation between cognitive science and the philosophy of thought experiments. Specifically, as cognitive science can be used to support contradictory claims in philosophy, I conclude that being more careful about how we frame the paradox of thought experiments will save precious ink (or computer pixels). One thing we learn from cognitive science that cannot be denied is that the aims and means of the mind are far more varied than traditional epistemology has assumed. I take this lesson seriously in seeking an answerable version of the paradox in Chapter 5.
Chapter 4
Thought Experiments and Cognitive Science

Accepting hypotheses merely on the basis of thinking about them constitutes a kind of epistemic hubris.

(Thagard 2014, 288)

Paul Thagard is a well-known cognitive scientist and philosopher of mind who has recently expressed skepticism about the cognitive efficacy of thought experiments. In so doing he joins forces with Meinong (1907), Duhem (1954, 201-205), Dennett (1984), Dancy (1985), Harman (1986) and Wilkes (1988). According to Meinong, who was perhaps the first skeptic about thought experiments explicitly so-called, “an experiment that in fact does not exist at all, can neither prove nor teach anything” (1907, 276-277). Dennett, Dancy, Harman and Wilkes are no less forceful. What sets Thagard apart is the source of his skepticism: cognitive science. This is noteworthy because cognitive science is heavily drawn upon by the largest and fastest growing group of people writing about thought experiments today, for example, Bishop (1998), Cooper (2005), Gendler (2004), Palmieri (2003), Nersessian (1992, 1993, 2007), McMullin (1985), and Miščević (1992, 2007). This is why it is so important to address Thagard’s skepticism: before we accept a cognitive science-based account of how thought experiments work, we must first be sure that cognitive science does not also show that thought experiments are not to be trusted.

Briefly, Thagard claims that “the made-up thought experiments favoured by many philosophers are not evidence at all” (2010a, 209). Indeed, “philosophical attempts to establish truths by a priori reasoning, thought experiments, or conceptual analysis have been no more successful than faith-based thinking has been. All these methods serve merely to reinforce existing prejudices” (2010a, 41).

---

4 An earlier version of this chapter has been published as Stuart 2014.
This chapter provides reasons to reject Thagard’s skepticism. I will begin with an outline of his phenomenology of thought experiments, and then address his case for the view that thought experiments are dangerous and misleading when they are used as evidence. While I reject his view about thought experiments, I accept the cognitive science he draws upon. In fact, I will show that the same studies can be used to make a more positive contribution to the field. This chapter is therefore an effort to convert Thagard and anyone tempted to agree with him, but also to present some new directions for research on the intersection between thought experiments and cognitive science.

1 Thagard’s Phenomenology of Thought Experiments

Thought experiments for Thagard are “mental constructions of an imaginary situation in the absence of attempts to make observations of the world” (2010a, 254). This independence from observation makes their output a priori. Thagard defines a priori knowledge as “knowledge that is gained by reason alone, independently of sensory experience” (2010a, 251).

According to Thagard, a priori methods in general, and thought experiments in particular, function by eliciting “innate ideas, concepts that we have at birth” (2010a, 36). This is because thought experiments—especially those in philosophy—work by eliciting intuitions about a concept. What that concept is and what we can conclude about it by this method are results only of reflection, not careful scientific investigation via experiment. If we occasionally privilege thought over laboratory experiments, it is because we believe there are some concepts that can be investigated only by non-empirical means. The only concepts thus accessible (according to Thagard) are innate ones. When we ask ourselves what we would say about people who could split like amoebas (as in Parfit 1984), we are performing conceptual analysis on some non-empirical concept of personal identity. If mere reflection yields knowledge about such a concept in a way that does not compete with scientific investigation, that concept must have been innate. Such concepts have been at the center of philosophical debates for centuries, but due to the advent of philosophical naturalism and the growing applicability of cognitive science to philosophical questions (Thagard 2010a, 6), it is time to leave these concepts behind for more empirically grounded ones.
Why then have thought experiments persisted in philosophy and science? Thagard answers that many philosophers are still blinded by the promise of *necessary truth*. While science only offers contingencies, thought experiments explore the nature of concepts like KNOWLEDGE, REPRESENTATION, REFERENCE, IDENTITY, and so on, and what they discover will be true without the possibility of empirical falsification.

According to Thagard this promise is precisely what makes thought experiments so misleading. If the goal of thought experiments is to find necessary truths, then they do not serve their purpose, since “the notion of necessary truths is just as empty as the notion of the a priori” (2010a, 38). Believing the opposite leads only to wasted time.

But surely not all thought experiments function this way; many are employed by well-known scientists and can be found in science textbooks. For a naturalist like Thagard, this responsible use of thought experiments must be differentiated from the others. Thagard does this by claiming that *no* thought experiment we find in a science textbook is meant to *justify* a scientific proposition. Thagard knows of “no case in science where a theory was adopted merely on the basis of thought experiments” (2010a, 39). For Thagard, thought experiments in science only play a role in the context of discovery and never in the context of justification. Thought experiments might be useful for “suggesting and clarifying hypotheses” (2010a, 60) and for “revealing inconsistencies in opposing views” (2010a, 39). But under no circumstance should we try to use them to justify a belief (2010a, 60).

Thagard’s criticism does not stop here, perhaps because he realizes that many thought experiments focus on abstract but still empirically relevant concepts. After thought experiments that investigate innate concepts and those that are used responsibly by scientists to “pump intuitions” (Dennett 1984, 2013), these form a third class of thought experiments that Thagard must address. While thought experiments that focus on innate concepts do little more than waste the time of philosophers, thought experiments meant to justify empirical claims this way can lead to empirically false and therefore dangerous conclusions.

To summarize, there are three sorts of thought experiment, for Thagard: 1) those that investigate innate concepts that are unconnected to experience, 2) those that mistakenly investigate empirical concepts as though they were innate, and 3) those that pump intuitions in science and
generate useful hypotheses or check consistency. I should note that charity demands the distinction between the first and second kind of thought experiments for Thagard because without it we come to a contradiction. On the one hand, Thagard claims that a priori propositions are not to be trusted because they cannot be confirmed or disconfirmed by experience. They are totally disconnected from the world. On the other hand, he claims that a priori propositions are to be avoided because they are often false, and this has been shown empirically. He cites our intuitions about Euclidean geometry as an example (2010a, 36-37). These two claims cannot both be true, unless they refer to mutually exclusive sets of a priori propositions. Therefore we must assume that Thagard is criticizing different sets of thought experiments instead of different features of the same set of thought experiments. This does not affect his main argument; which in his (2010a) is principally one of analogy between thought experiments and religious reasoning.

Believers of all sorts use religious reasoning such as prayer and meditation in an attempt to find truth about non-empirical religious matters, and this corresponds to the first use of thought experiments as a means of probing innate ideas. But they may also try to use such reasoning to investigate empirical matters, which corresponds to the second improper use of thought experiments. These are essentially the two problems with faith that Thagard identifies in thought experiments. In more detail, here is the first:

A priori reasoning...has the same arbitrary nature as faith. Just as what people adopt as their particular brands of religious faith depends largely on their upbringings and associates, so what people take to be true a priori—what they can imagine—depends on what they have already learned. (2010a, 37)

What someone finds intuitive is a function of their life experiences. What a mathematician finds obvious about identity will probably be different from what a plumber or a priest finds obvious. Since our intuitions are partially the result of chance events that helped to form our system of beliefs, there can be no objective method of adjudication when we disagree about what is intuitive. We cannot say that someone who has a different intuition about an imaginary case is wrong.

Now, focusing on the empirical disconnectedness of a priori reasoning, we see that
Without...evidential evaluation, use of thought experiments becomes merely the trumpeting of one philosopher’s intuitions over another’s, a process no more conducive to truth than the professions of faith by advocates of rival religious sects. For every thought experiment there is an equal and opposite thought experiment. (2010a, 39)

That is, thought experiments are more like subjective reports of opinion than a means of access to the world. The imaginary scenarios we invent do nothing more than provide rhetorical banners behind which we pronounce our beliefs. The banners themselves justify nothing.

The second problem applies to the empirical use of a priori reasoning, and it was alluded to above: faith and thought experiments lead too often to false beliefs, and so they are not empirically reliable. The strength of this criticism depends on case studies, and Thagard provides many examples of thought experiments that he believes have led science astray, like Plato’s cave and Searle’s Chinese room (see Thagard 2014).

Thagard therefore sees two different sorts of problematic thought experiments (innate and empirical), each with different defects (useless and false). Having drawn this distinction, I will leave Thagard’s comparison of thought experiments to religion behind. His points about thought experiments could be correct independently of whether they also apply to religious beliefs.

2 Against Thagard’s Phenomenology of Thought Experiments

Before replying to Thagard’s skepticism, there is an important caveat to be made. His points about the evidential power of thought experiments are meant to be limited to philosophical and not scientific thought experiments because the former are used to justify claims, while the latter are properly limited to the generation of new hypotheses and finding inconsistencies. However, Thagard’s naturalism implies that the methods of philosophy and the methods of science are (or should be) one and the same. Presumably then, Thagard would allow that thought experiments can be used responsibly in philosophy as well as science, to point out contradictions and inspire new ideas. The real issue is that some thought experiments are believed to yield evidence for claims about the empirical world, and for Thagard this is never justifiable. Therefore in the
remainder of the paper I will not differentiate between thought experiments in science and philosophy, but rather focus on whether thought experiments can provide evidence.

In this section I respond first to Thagard’s notion of a priori reasoning and why he thinks it should not be admitted into science, and then to his characterization of thought experiments and their role in the progress of human inquiry.

2.1 Thought Experiments as Independent of Experience

As we saw above, thought experiments are a priori for Thagard. By “a priori” Thagard simply means “independent of experience.” But there are several ways a proposition can depend on experience. Here are three. 1) A proposition might depend on experience in the sense that some experience is necessary in order to understand the proposition. 2) A proposition might depend on experience in the sense that some experience is necessary to cause us to entertain that proposition. And 3), a proposition might depend on experience in the sense that some experience is necessary in order to justify that proposition (Boghossian and Peacocke 2000, 1-2).

Of interest in the present context is the third sense of dependence. That is to say, a proposition can depend on experience in the first two senses, and still be a priori. Here is how. Consider these two sentences: “The green car is green” and “the green car is mine.” The first can be known a priori, even though we would not be able to understand this statement without some general experience, for example, with colours. This is the first sense of dependence. The second sentence is different. While it does rely on previous experience for comprehension, it also calls for some specific experience to justify our knowledge of it (perhaps the memory of signing ownership papers). This is dependence in the third and relevant sense.

Turning to the second sense of dependence, suppose we come to know that modus ponens is valid after seeing a proof written out in symbols on a piece of paper. Those symbols cause us to entertain a proposition, and we need various perceptual capacities for this to happen. This is the second sense of dependence. But perceptual experience is not what justifies the conclusion. That is done by the proof itself (Peacocke 2000, 255). This type of knowledge is independent of experience in the third, relevant sense.
Thagard does not distinguish between these senses. He assumes that dependence in any of them makes a proposition or method a priori. In this way his conception of the a priori is conceptually underdetermined and leads him to claim that a priori beliefs must be based on innate ideas. Perhaps he is right that anything independent of experience in all three senses would have to be innate, but there are good reasons to think that these senses can and should be separated. Therefore, thought experiments can be linked to experience in important ways, even if Thagard is correct to characterize them as a priori.

However, if we suppose that a priori reasoning really is restricted to innate ideas, we could still perhaps show that they are connected to experience in an appropriate way. If innate belief-structures are the product of evolution, they may well have originated from human experience, although indirectly through the operation of natural selection on the belief structures of generations of our ancestors (as argued in Sorenson 1992, 51-75). This means that even if thought experiments are a priori and function by the use of innate ideas, they need not be independent of experience in a way that disconnects them from reality.

I conclude that Thagard’s characterization of the a priori is too crude to be of any help in assessing the evidential significance of thought experiments. “Independent of experience” is a criterion that can be satisfied without the need for innate ideas. And once we see that empirical content can be involved in a priori reasoning, it begins to look more plausible that such reasoning could stand as evidence for claims about the world.

2.2 Thought Experiments as a Means to Necessary Truth

There are several problems with Thagard’s claim that thought experiments are intended to lead to necessary truth. First, this claim imputes a very specific goal into the heads of thought experimenters everywhere. This is problematic. Second, it assumes that some of the philosophers who discuss thought experiments believe that this is the aim of the method, and I think that is false.

There are figures in the history of philosophy who might have agreed with Thagard that thought experiments could reveal necessary truths. Perhaps Kant is an example. But the majority of contemporary philosophers do not see this as one of their primary functions. For example, John
Norton argues that thought experiments are nothing but arguments. While some arguments, for example, mathematical ones, might attempt to yield necessary truth, it is certainly false to say that all do this, especially inductive and informal arguments. Presumably, Thagard will not object to thought experiments if they are merely arguments.

Besides Norton, there are philosophers who believe that thought experiments are manipulations of mental models. The claim here is that thought experiments create a “representation of the problem, i.e., [a] representation giving the opportunity for the subject to manipulate the problem situation (in his head of course) in a particularly easy fashion, and so that it makes it easy to mobilize [the] subject’s cognitive resources—skills, implicit background knowledge, perceptual beliefs, etc., in a way superior to regimented reasoning” (Miščević 1992, 224). In other words, thought experiments are actions that connect and manipulate representations in a way that brings us to something new. Their output is only as reliable as the representations they manipulate, and the cognitive mechanisms they use. Since our representations and cognitive mechanisms are imperfect, we do not expect necessary truth from thought experiments thus conceived. These philosophers are broadly speaking naturalists like Thagard, and I will argue below that Thagard should ally himself with their cause.

But there are also rationalists in the debate, and perhaps Thagard is concerned with these philosophers who claim that the operation of reason alone can produce new knowledge. There is even a Platonist who argues that thought experiments can provide a glimpse into Plato’s heaven. In the past, there have been philosophers who believed that whatever is revealed or deduced by the power of reason is necessarily certain. But Thagard’s opponent is not Plato, Descartes or Kant. We should note that almost all modern day rationalists, including Bealer (2002, 74), BonJour (1998, 110-115), Friedman (2001, 83-103), and Peacocke (2000, 257), accept fallibilism about a priori knowledge. This includes Platonists and non-Platonists who believe that thought experiments are a priori, including Arthur (1999, 227) and Brown (1986, 6, 8, 10; 1991, 44, 99; and 2004, 31-32).

Having quickly surveyed the broad position types found in the literature, it is not clear that there is anyone who holds the thesis that thought experiments are meant to reveal necessary truths. If
no one holds this thesis (and Thagard does not name names), then Thagard is attacking a strawperson. But I would like to say something on behalf of the strawperson anyway.

How might thought experiments lead to necessary truth in an epistemically responsible way? One way is for thought experiments to tell us something about the essence of a concept. When Thagard discusses thought experiments as a means of doing conceptual analysis, we are given examples that fit this characterization, that is, we are given thought experiments that seek necessary and sufficient conditions for our concepts. If the concept in question is something like ELECTRON, then I believe Thagard is right to be skeptical. We should not pretend that mere thought could reveal the necessary or sufficient conditions for membership in this concept’s extension. However, it is another matter when we are dealing with concepts that are more or less of our own construction. There are concepts which are more defined than discovered, and the necessity that results from a thought experiment about these concepts can be conceptual rather than metaphysical. It is a mistake therefore to claim that all thought experiments that function by conceptual analysis aim at necessary truth and are epistemically irresponsible for this reason, because conceptual necessity is not dangerous. It is often nothing more than the establishment of a useful tautology. Now if Thagard allows for thought experiments dealing in conceptual necessity, perhaps because these help in revising our theories and developing hypotheses, then he should not denounce thought experiments for their kinship with conceptual analysis.

This is not to say conceptual analysis is without its problems. As mentioned, I share Thagard’s skepticism concerning instances in which necessary truths about the physical world are purportedly established by nothing more than mere thought. But such cases are the exception: it is rare to find a thought experiment that claims to have established a fact about the physical world (or any empirical concept) without relying on supplementary evidence. Thought experiments provide varying degrees of support; from suggestion to establishment. And we should not make the mistake of thinking that because a thought experiment deals with something that is necessary if it is true (for example, \(x\) and \(y\) are related by law of nature \(z\)), the thought experiment itself is meant to be the sole ground of that necessity. Evidence for a claim that is thought to be metaphysically necessary can be weak. Einstein cannot run as fast as a wave of light in his mind because this would force the wave to become static and no longer what it is, an oscillating electromagnetic wave (Einstein 1949a). This thought experiment can be taken as
evidence that there is something special about electromagnetic waves that limits our ability to change their apparent motion by changing our own (see Norton 2012). Obviously Einstein’s thought experiment had to be supported theoretically and empirically before it was accepted by the scientific community, but the thought experiment still provided prima facie evidence for a claim, and prima facie evidence is evidence. The same can be said about a priori reasoning in philosophy, for example, concerning Thomson’s violinist thought experiment that supports a woman’s right to abortion. If we are brought to believe that we can remove a person who has been connected to us without consent, this is prima facie evidence that the right to life is separate from the right to bodily integrity. That is, abortion is in some cases primarily an exercise of the right to bodily integrity, and only secondarily a displacement of the right to life (Thomson 1971). Thus, a priori reasoning can be used to make a case for a metaphysically necessary proposition, but there is no reason to claim that it must establish such a necessity all on its own in an epistemologically irresponsible way.

This section has addressed some of Thagard’s arguments against the first class of thought experiments—those that are criticized for being isolated from experience and drawing only on subjective opinions about innate ideas in order to yield necessary truth. I concluded that thought experiments are not isolated from experience, even if they are a priori. And they need not operate solely using innate ideas, although even if they did, this might not isolate them from experience. And finally, thought experiments do not always aim at necessary truth, although when they do, they are not always objectionable for this reason, since there are different kinds of necessity we might pursue.

2.3 Against Thagard’s Definition of Thought Experiments

Thagard’s skepticism about thought experiments is guided by a poor definition of thought experiments. He defines them as “mental constructions of an imaginary situation in the absence of attempts to make observations of the world” (2010a, 254). For one, this definition ignores the performative aspect of a thought experiment. In thought experiments we do not merely construct an imaginary situation, we do something with it.

Second, Thagard’s definition implies that all thought experiments must remain physically unrealized for them to count as thought experiments. But surely if we perform Galileo’s thought
experiment about falling bodies, this does not make Galileo’s thought experiment stop being a thought experiment. And Thagard himself notes that this “thought experiment” has been performed (2010a, 25-38), so I believe that Thagard implicitly realizes that attempting to make an observation is not especially relevant for what makes something a thought experiment.

Finally, his definition fails to distinguish between thought experiments and other kinds of imagined scenarios or works of fiction. Whatever the relation between fiction and thought experiments is, it will do no good to claim that every imagined scenario is a thought experiment. Such a definition misses what is interesting about thought experiments, namely, that they aim to improve our epistemic position. Staring up at the clouds and imagining creatures locked in mortal combat must be distinguished from the practice of thought experimentation, if only for practical epistemological purposes. The Einstein-Podolsky-Rosen thought experiment is not in the same business as cloud-gazing.

To summarize, Thagard's definition of thought experiments will not do. It would be better to adopt one of the more standard definitions in the literature. This would not substantially affect his arguments against thought experiments, because what he objects to is not the mere formation of mental constructions in either science or everyday life, but placing evidential weight on the intuitions elicited by manipulations of such constructions. Let us then see what can be said against his specific objections to the epistemic use of thought experiments.

2.4 The Method of Thought Experiments

Above, I made the case that thought experiments do not usually aim to establish necessary truths. So what else might they do, and how? Brown suggests at least the following roles for thought experiments: they may test scientific conjectures (by refuting or confirming them); illustrate theories; simulate or uncover natural phenomena, and create new phenomena entirely (1991). Roy Sorenson claims that thought experiments are mainly used to test the modal status of propositions, and “the favourite use of thought experiment is to establish a possibility” (1992, 36). Sorenson adds that thought experiments may also be used to control extraneous variables, invert ideals of natural order, and serve as a “master test for conceptual analysis” (1992, 11-16). John Norton sees thought experiments as expanding our knowledge by induction and deduction.
(1996). Catherine Elgin claims that they “exemplify” properties and relations in a way that makes interesting or relevant features of the world stand out (Elgin 2014).

Thagard disagrees. He sees only two main modus operandi for thought experiments, corresponding to the two improper uses of thought experiments that I distinguished at the beginning: the eliciting and broadcasting of arbitrary intuitions, and the inference from what we can imagine to what the world must be like. We have already discussed the first, so I will now turn to the second. My response is that not all thought experiments use this inference, and those that do are not necessarily unreliable for this reason.

2.5 From Conceivability to Possibility

In Galileo’s falling bodies thought experiment, we do not ask ourselves whether the situation presented is conceivable or possible, because we could easily make it actual. Its conceivability might be a necessary precondition of our performing the thought experiment, but on its own, it is irrelevant for the justification of the thought experiment’s conclusion. One thing that helps to justify the thought experiment is that it predicts what would actually happen if it were performed, ignoring considerations like wind resistance. Another thing that justifies it are its empirically well-supported premises, and the fact that it employs only assumptions that its opponents would accept (Popper 1959)

A more complex example is Schrödinger’s Cat (Schrödinger 1935). We might think that the point of the thought experiment is to see that it is possible for a cat to be neither dead nor alive nor neither nor both, given that we can conceive of it as connected to the type of quantum system Schrödinger describes. But in fact, we cannot properly conceive of the cat in superposition, just as we cannot conceive of a particle in superposition. Equally, it does not really matter if such a set-up is possible: it looks like macroscopic superposition might be actual (Romero-Isart et al. 2010). The point of the thought experiment is rather to show that when we combine the Copenhagen interpretation (which takes the superposition of an object to be the occupation of all possible states of the observable variable until observed), with the Schrödinger equation (which implies that any number of systems may be “entangled”), then we are forced to admit that macroscopic systems like cats must be capable of “resolving” in the same way that wave packets do. This is not an exercise in innate ideas. It does not point out a contradiction or generate a new
hypothesis, and it does not lead us astray by concluding that something is possible because we can imagine it. Instead, the thought experiment seems to rely on perfectly standard types of inferences. It forces physicists into taking a stand on exactly when we should say that the cat has survived or perished, and what role observers play in the “collapse” of entangled quantum systems. If the cat only “resolves” when an observer opens the box, then what does the cat see while it waits?

There are many other thought experiments that are empirically relevant whose justification does not rely on the conceivability-possibility inference (for example, Poincaré’s Disk, Einstein’s Elevator, EPR, etc.), but I would like to address now those that do. In the last two decades, much work has been done on the inference from conceivability to possibility, and it is generally agreed that the former is at most only a guide to the latter. The relation is not and has never been one of entailment. Its purpose is to engage our modal faculty and enable fruitful modal discourse (see Gendler and Hawthorne 2002, Yablo 1993). Thought experiments that use this inference should therefore be understood as attempting to provide only a weak form of evidence for a modal claim, although evidence nonetheless.

Without trudging too deep into the murky waters of modality, it does seem that we must have at least some fallible way of knowing what is possible. This ability, however it works, is responsible for our capacity to reason counterfactually, which seems to be a necessary precondition of scientific thought in general (Buzzoni 2008, 116-120). If we could not imagine different ways the world might be, we could not find out by experiment what it is really like. For instance, we could not envisage instruments for testing possible values of a variable, or experimental setups that make such measurements possible. As long as conceivability is used as an aid to this faculty, and as long as we assume that methods of science are the right methods (which Thagard does), it should not be labeled dangerous or misleading.

Up to this point I have defended a priori reasoning and thought experiments against Thagard’s skepticism. Now I will show that these tools of reasoning are reliable and perhaps even necessary for doing good science according to Thagard’s own conception of “good” science.
2.6 Thought Experiments as Weakly Evidential

For Thagard, the archetype of good reasoning is scientific reasoning. He claims that we can delimit good from bad science as follows:

In general, we can use descriptive information to help generate normative conclusions whenever we can identify the appropriate goals. If the appropriate goals of science are truth, explanation, and prediction; and if the history of science reveals that experiments and inference to the best explanation are the best practices for achieving these goals; then these practices are normatively justified as what scientists ought to do. (2010a, 175)

Thagard argues that inference to the best explanation and experiment are the practices that achieve the goals of science, so they are normatively justified in science and presumably any other field with similar goals. But then the practices that make inference to the best explanation and experiment possible would also (indirectly) lead to truth, explanation and prediction. Is there anything that makes inference to the best explanation possible? Thagard claims that we find the best scientific explanation by seeking coherence and avoiding inconsistency (2010a, 21-22). And these ancillary goals, Thagard elsewhere admits, can be achieved using thought experiments. He says, “thought experiments are fine for suggesting and clarifying hypotheses” (2010a, 60), and also for “revealing inconsistencies in opposing views” (2010a, 39). Thus, thought experiments can (at least indirectly) help us to reach truth, explain phenomena, and make predictions. Insofar as thought experiments do this, they provide evidence for claims, explanations, or predictions. This shows that Thagard is mistaken to claim that thought experiments can or should play no evidential role in science. According to his own definition of good science, and his own characterization of the roles that thought experiments can play, thought experiments are a useful evidential tool for the scientist.

2.7 Thought Experiments as Strongly Evidential

If there is an experimental side to thought experiments then they are evidential, since Thagard’s definition of “evidence” highlights information gained by experiment (see below). I think I can show that there is such a side to thought experiments, first by expanding Thagard’s definition of experiment, and then tightening it. Thagard defines experiments as “planned manipulations” that
alter only a few features of a system to “be able to identify causes and effects.” They are “repeatable,” and they make possible precise quantitative measurements (2010a, 25). He is right that thought experiments are not experimental under this definition. For one, they do not usually make precise quantitative measurements possible. However, I would point out that many accepted physical experiments also lack this quality—for example, the celebrated experiments conducted by Pasteur that have been taken to disprove the thesis of spontaneous generation (see Conant 1953). These experiments showed that life will not spring from inanimate material such as dead meat if we remove the presence of air containing “spores” or “germs.” Fleas and maggots and mice come from other fleas, maggots and mice, not from dust or dead meat or dirty rags. Under the condition that all experiments must make possible precise quantitative measurements, experiments like those of Pasteur are disqualified.

Leaving this criterion out of Thagard’s characterization, we must still face the others, and it is true that many thought experiments do not identify causes or effects. However, once again, many laboratory experiments fare no better. For example, Michelson and Morley’s famous experiment (1887) suggested only that the ether theory was empirically not well-supported; it suggested no cause or effect. Examples of this kind can be multiplied indefinitely by considering all other important experiments with negative conclusions.

We are left with repeatability, which is a criterion that thought experiments are able to satisfy. Even if we envisage the requisite imaginary scenarios in a thought experiment slightly differently, this would not prevent that thought experiment being successfully repeated. Whether we imagine Galileo’s cannon and musket balls to be black or grey does not matter, and no one would complain that they accidentally performed a different thought experiment if they were informed that in Galileo’s mind the falling objects were originally purple. If we cannot discuss our own performances of a given thought experiment as performances of the same experiment, then it is only because human communication is impossible in general.

Could Thagard amend his definition of experiment to exclude all thought experiments and retain all laboratory experiments? I think this would be very difficult. “Experiment” is a general term for good reason, which makes it more and not less likely that thought experiments will count as experimental.
Of course, we should also concern ourselves with the characteristics that do unite the practices that fall under the spectrum of applicability of “experiment.” Let us first admit that actions can be more or less experimental. Also, all experimental activity seems to test the change in some few variables while controlling others. Finally, it always takes place in response to the formulation of a hypothesis, and it relies on idealization and abstraction. In a computer simulation, we manipulate an idealized representation of a system with the hope that its output will be reliable with respect to that system. This is at least somewhat experimental (see Morgan 2003, 2005; Guala 2002; Parker 2009; Morrison 2009; Giere 2009; Lenhard 2007; and Barberousse et al. 2009).

Once we acknowledge the seemingly experimental properties of computer modelling, we should note that it is a short step from a computer model to a mental model or thought experiment (Di Paolo et al. 2000). Each of these methods manipulate representations using more or less specific inference patterns, both reflect heavily the theory from which they are developed, and both rely on structural analogies between target and represented system, as opposed to the physical analogies exploited in laboratory experiments. My point in all this is that Thagard must provide a principled way to set apart computer simulations and thought experiments epistemically, or else he will be forced either to include both as evidence-producing tools of scientific reasoning, or let go of computer simulations, which have formed a major part of his work since the 1980s.

### 2.8 Evidence

Finally, what exactly does Thagard mean when he says that thought experiments cannot count as evidence? Popular notions of evidence are quite divergent (see Chapter 1). Thagard’s own definition of evidence is: “information gathered by careful observation, especially through scientific experiments” (2010a, 252). Can thought experiments produce information gathered by careful observation? That depends on what we count as careful observation. If Thagard means to exclude introspection, then we lose much of our evidence for claims like “I am hungry,” or “I am upset.” On the other hand, there are many cases where “observation” seems inappropriate to characterize an instance of evidence. For example, we receive evidence from arguments, computer simulations, mathematics, climate models, and many other methods which function by manipulating representations within predetermined system constraints or transforming
propositional content according to truth-functional rules. None of these function by observation. The output of an argument, simulation, proof, model or thought experiment is trivially observable if we write it down or have it presented on a screen, but that is not what justifies its output. Evidence is a normative epistemic concept related to justification, and observation is too narrow to give us that normativity. Again we find ourselves in a situation analogous to the one above: either we constrain evidence to information gained by observation in a sense that excludes thought experiments and lose the evidential capacity of other means that we normally think of as evidential, or we recognize that the notion of evidence is flexible for good reason, which again makes it more and not less likely that thought experiments will be able to serve as evidence.

It has been my aim to show that none of Thagard’s arguments against the evidential significance of thought experiments are decisive. In what follows I consider some of the results from cognitive science on which Thagard draws in order to explore the ability of thought experiments to provide evidence.

3 Cognitive Science, Thought Experiments and Mental Models

With Terrence Stewart, Thagard has presented a new account of human creativity (2011). It is interesting to note that in this article the authors cite Nancy Nersessian’s account of mental models approvingly. Nersessian defines a mental model as a “structural, behavioral, or functional analog representation of a real-world or imaginary situation, event or process” (2008, 93). Thagard and Stewart go on:

We agree with her contention that many kinds of internal and external representations are used during scientific reasoning, including ones tightly coupled to embodied perceptions and actions…our account is certainly compatible with Nersessian’s (2008, 135) claim that conceptual changes arise from interactive processes of constructing, manipulating, evaluating, and revising models. (2011, 25)

This is telling because Nersessian is well-known for an account that explains the cognitive power of thought experiments by identifying them with the manipulation of mental models (1992, 1993,
and 2007). Thagard is willing to accept that mental models can produce new knowledge, and specifically that it is through their use that creative new conceptual combinations are made possible. To explain how we gain understanding from thought experiments in a way that is consistent with Thagard’s work in cognitive science, I will assume the characterization of thought experiments as mental models, which is legitimate because even opponents of this view only argue that this characterization does not on its own explain the epistemological efficacy of thought experiments. For instance, Brown (2007c) and Norton (2004b) both allow that this characterization might be descriptively accurate, yet nevertheless consistent with the truth of their accounts since manipulating a mental model can bring us to see Platonic truths or instantiate an argument form. To explore this characterization, I will first present Thagard’s understanding of mental models and how we learn from them.

For Thagard, mental models are representations, “consisting of patterns of activation in populations of neurons” (2010b, 447; see also 2010a, 78). A neural population is “a collection of neurons that are richly interconnected,” and a population of neurons “represents something by its pattern of firing” (2010b, 450). Mental models and concepts are often “produced by, and maintain some of the structure of, perceptual inputs” (2010a, 78), and presumably, the way they maintain the structure of their objects is what makes them reliable. The idea is that “a pattern of activation in the brain constitutes a representation of something when there is a stable causal correlation between the firing of neurons in a population and the thing that is represented, such as an object or group of objects in the world” (2010b, 450). That is, neural populations are causally connected to the things they represent because they preserve structure. If I see something to the left of something else, and I represent that situation to myself, my representation of one object will be in a neural population that is physically (or if you like, neurally) to the left of the other, and this will instantiate my concept “to the left of.” The same goes for temporal and other kinds of structures that may be preserved in thought.

We dynamically manipulate these neural structures to create mental models. Mental models are a product of the interaction of higher-level mental representations like “cause,” which are themselves products of lower-level representations stemming from the senses or emotional faculties. We know that it is higher-level interactions that produce the mental model because mental models are conscious, and lower-level interactions are not. What Thagard calls “top-
down” (mind-to-neuron) processes monitor and guide the way the model is run. These top-down processes work with the bottom-up processes, such as simple representation, and they all run together, creating feedback loops which ensure that what is represented and manipulated maintains its structural similarities to the target system.

Mental models contribute to scientific discovery, according to Thagard, by enabling the generation of genuinely new concepts, such as SOUND WAVE and WIRELESS EMAIL (2010b, 3-4). What happens is that any number of conceptual, linguistic, perceptual, or emotional representations may be combined in a “kind of twisting together of existing representations.” This Thagard calls “convolution” (Thagard and Stewart 2011, 2). The mathematical function from which this process gets its name takes several functions as its domain, and yields a blended overlap as its output. Thagard wishes to characterize the combination of concepts in terms of this function, and to do this he shows that we can make a realistic computer model of human creativity using it. Our representation process begins by converting sensory input (for example, light and sound waves) into electrical impulses in the brain. The computer simulation uses groups of “nodes” which are designed to mimic the stochastic nature of real neurons. These represent different types of input into the language of vectors by coding it using a series of “natural transformations” (2011, 11). He can then make the model “represent” and convolute sensory input, as in the following figure:

![Fig. 6: “Convolution Occurring In Simulated Neurons” (Thagard and Stewart 2011, 15)](image)

Suppose that the computer model is presented with the two visual stimuli on the left. These will be converted into vectors and represented by a network of nodes. The representations are then convoluted to create the new “neural” network on the right, which can be translated into a new visual output (far right). Thagard points out that “the convolution of the two input patterns is not
simply the sum of those patterns and, therefore, amounts to an emergent binding of them” (2011, 15). This process reliably produces new content, which can become a candidate for knowledge.

*Creative* convolutions are those that are accompanied by the “Aha!” feeling we get when we suddenly realize something important. This response is an emotional reaction triggered by the experience of coherence between what a thinker wants or needs, and what the new combination makes possible. This emotional response becomes part of the convolution which includes the new concept or idea. Thagard and Stewart sum it up this way: “The Aha! experience is not just a side effect of creative thinking, but rather a central aspect of identifying those convolutions that are potentially creative” (2011, 19).

With the rudiments of this account now in place, I want to sketch how it interacts with other cognitive science-based accounts of thought experiments.

Thagard’s description of mental modelling can be used to substantiate Nersessian’s account of thought experiments. She writes,

> When a thought experiment is successful, it can provide novel empirical data. These data are novel in the sense that although they are contained in current representations, the means to access them were not available until someone figured out how to conduct the thought experiment. (2007, 127)

Thagard’s work substantiates the notion of “opening a means of access.” A mental model brings us to a new conceptual combination by connecting patterns of activation in different neural populations. The information is already contained in the current representations, and it is made available when the patterns are convoluted into something useful. This is brought to our attention when an emotional reaction promotes the new datum. In some sense a house is *in* the assorted building materials when they arrive to a construction site, but its final functional form is something new, something that was not there before, which is made available by a specific process of combination (the psychological realization of which would naturally produce an emotional reaction).

Thagard’s research also strongly suggests that mental modelling involves a great deal of non-propositional content. It is multi-modal; that is, it involves input from emotions, linguistic
concepts, physiological states and representations. Those who adopt the mental models account of thought experiments agree that the mechanisms capable of bringing us new knowledge, and the unarticulated nature of the resources upon which they draw, are often non-propositional (see Miščević 1992, 222-223, 225; 2007, 189; and Nersessian 1992, 292-293, 297). Thagard’s multi-modal theory of mental cognition provides additional evidence for this claim. However, Thagard’s account takes this idea even further by specifying the nature of the non-propositional elements: they are non-propositional because they are activities of neural populations that refer using something like structural isomorphism.

Miščević states that mental models seem to have a quasi-spatial character. If Thagard is right, the spatial and temporal character is actually full-blown, since the brain retains both relations in a literal sense. Miščević is also impressed by the speed and ease with which thought experiments bring us to a conclusion (Miščević 2007, 200). According to Thagard’s account, much of what happens in a thought experiment is done unconsciously. By the time the aha! moment arrives, there has already been a great deal of unconscious convolution, which helps to explain how a process so fast could also be reliable.

4 Conclusion

Thagard claims that thought experiments are dangerous and misleading when used as evidence. He understands them as an a priori method aimed at necessary truth. He accuses them of operating in isolation of careful empirical observation and relying only on subjective intuitions. They are therefore nothing like real experiments. I have argued that all of these claims are either false or based on false assumptions. Thought experiments are an exciting part of the scientific method, and philosophers of science must try to understand them. This should be a goal Thagard shares, and so I closed with a short discussion of a few of the interesting directions that his own work in cognitive science suggests.

That being said, I think Thagard, like other skeptics about thought experiments, is right in his motivation. Those who explain how thought experiments work do indeed take it for granted that thought experiments provide something of value in science and elsewhere. This assumption must be grounded. Thagard’s challenge is perhaps overly harsh, but the project is an important one.
The next chapter takes this question seriously in an effort to get clear on what exactly (an interesting set of) thought experiments do in science. Only after such an exercise can we consider the applicability of cognitive science. Otherwise, we run the risk of trying to justify a role for thought experiments that they do not play. I am adopting here the Baconian strategy mentioned in Chapter 1, which requires that we start with a study of what we do when we experiment in thought.

Let me repeat in order to make the connection between this half of the dissertation and the next as clear as possible. Norton presents an account to answer what he calls the epistemological question of thought experiments, which characterizes thought experiments as attempting to generate new knowledge. I evaluated this account and found it unsatisfactory. Thagard suggests another set of epistemological roles for thought experiments, which I also found unsatisfactory. But there are other roles that thought experiments play. These are sometimes mentioned but not often investigated. For example Norton tells us that his concern “is not directly with the many other interesting facets of thought experiment: their effective exploitation of mental models and imagery, their power as rhetorical devices; their entanglement with prior conceptual systems; their similarity to real experiments; and so on. More precisely, [he is] concerned with these facts only insofar as they bear directly on the epistemological problem.” (2004b, 44-45).

The degree to which thought experiments are like laboratory experiments surely is relevant for the epistemological problem. The more thought experiments are like laboratory experiments, the more reason we have to think they can provide knowledge. And if thought experiments extend our knowledge by bearing on conceptual systems, that is also epistemologically relevant. Finally, if we reject the distinction between discovery and justification, we should also expect rhetorical devices to be important for the progress of science. I think it is time to take these other roles more seriously, which requires spinning the wheels of paradox represented in Figure 4. In Chapter 1, I characterized the paradox of thought experiments as a set of related problems. All are legitimate problems to pursue. Norton singles out one of them, Thagard another. The accounts provided to address these problems are not satisfying, and I take this as my motivation to leave them here. But instead of pursuing another interpretation of the paradox that appeals to me a priori, I will try to let the cases speak for themselves.
Starting in the next chapter, therefore, I will analyze a small set of thought experiments. This means that the methodology of my dissertation will shift from philosophical to historical, specifically taking the form of textual analysis. In Chapter 6, I continue to proceed by case study, but here the studies will be drawn from social and cognitive science. In Chapter 7, I return to philosophical argument in order to extract an account of thought experiments from the preceding.
Chapter 5
Five Thought Experiments Reconsidered

One only understands the things that one tames.

A fox (Saint-Exupéry 1943)

In this chapter I will look at some of the different epistemological roles played by five thought experiments in science. My operating assumption is that if we allow for there to be several epistemological roles for thought experiments in science, we complicate but precisify our understanding of them.

Thomas Kuhn was right about some thought experiments: they can help drive scientific progress by criticising the concepts of old scientific theories and/or introducing new ones (1977). The new concepts become part of a new theoretical background, which is important for the way we understand the world as described by science. Perhaps this is what Einstein does when he chases a lightwave in his mind, and Galileo when he imagines dropping bodies of different weights from a tower. Such thought experiments can indeed expose conceptual contradictions and suggest new concepts.

But the relationship between thought experiment and theory is not always one of criticizing or introducing concepts. A quick consideration of laboratory experiments will be instructive. In the past we might have said that the Michelson-Morley experiment justified the selection of Special Relativity over Maxwell’s theory of electromagnetic radiation. But thanks to developments in the history and sociology of science, we now know that the role played by laboratory experiments in scientific theory change is much more complicated (see, for example, Collins and Pinch 2012). The set of events referred to as “the” Michelson-Morley experiment contributed to Special Relativity in a generally positive way, but it was not until much later that it was portrayed as justificatory. Scientific experiments play many roles, some of which are rhetorical, and only one of which is the establishment of empirical fact. The same should be expected of thought experiments. For Kuhn, the pivotal question concerned scientific theory change. He recognized
both laboratory and thought experiments as part of this procedure, and he was right to do so. But even assuming that laboratory and thought experiments do sometimes play a justificatory role in scientific revolutions, we must be open to the possibility that they do more than just produce new facts, concepts, or contradictions.

A careful study of scientific thought experiments that appear during revolutions will show that there is a role for them which has not been paid sufficient attention in the philosophical literature. This is the production of scientific understanding as opposed to knowledge. Understanding is something of an umbrella term, which is advantageous in the present context because it can be used to cover several of the different epistemic goods produced by thought experiments that are not knowledge. Instead of going into definitional matters now, I will illustrate some of these epistemic goods using examples. (The relevant sense of understanding is expressed in Chapter 7, section 5). The reason I focus on thought experiments from revolutionary periods of science is because they are well-documented, and also the most discussed in the literature on thought experiments. I begin with two in-depth case studies—Maxwell’s demon and Einstein’s/Bohr’s clock in the box. These should illuminate sufficiently the role for thought experiments that I am interested in. I follow these with several shorter case studies—Einstein’s elevator, Darwin’s vertebrate eye, and Heisenberg’s microscope—to show that this account can be applied more generally.

1 Maxwell’s Demon

Maxwell begins publishing his now-famous papers on statistical thermodynamics in 1859. In 1860 he writes the “Illustrations of the Dynamical Theory of Gases,” which is a four-part paper, and “On the Dynamical Theory of Gases,” which contains the ideas that become the Maxwell-Boltzmann distribution. He also writes a paper in 1861 called “On the Results of Bernoulli’s Theory of Gases as Applied to their Internal Friction, their Diffusion, and their Conductivity for Heat.” A related paper is published in 1866. After this, not much more is written by Maxwell on statistical thermodynamics other than popular lectures, textbooks, and two short positive reviews—one for Willard Gibbs (1876) and another for his friend Peter Guthrie Tait (1878).

The theoretical structures (concepts, laws, principles) that Maxwell utilized in the papers from 1859-1861 were not new. Perhaps the two most difficult structures involved in the transition to
Maxwell-Boltzman statistical thermodynamics and Maxwell’s demon were entropy and the second law of thermodynamics. But the concept of entropy goes back at least to 1803 with Lazare Carnot, and was given full expression by Rudolph Clausius in 1850, whose version of entropy Maxwell straightforwardly adopts in *A Theory of Heat*. The second law was introduced by Sadi Carnot in 1824 (Carnot 1824, 55-59, 237-238) and brought into its modern form again by Clausius (Clausius 1856, 86. See also Klein 1970, 85-86). It is difficult to see how Maxwell’s demon could have played Kuhn’s role of concept introduction or contradiction if the demon does not introduce any new concepts or reject any old ones by pointing out inconsistencies. In my view, Maxwell’s demon therefore falls outside of Kuhn’s “important class of thought experiments.” In that case, what is it for?

It is first conceived in a letter to Tait, a supporter of Maxwell’s, in 1867. This comes seven years after the difficult groundwork of the new thermodynamics had been laid. It then appears in a letter to another friend, J. W. Strutt, three years later. It is finally published in *A Theory of Heat*, a popular expression of Maxwell’s ideas, in 1871. This is 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, v). Our knowledge has been extended, in the past tense, and the book will exhibit how. Specifically, Maxwell intends to show how several aspects of heat (and related phenomena) have already been explained and illuminated by experimental or theoretical developments in thermodynamics. It is possible that Maxwell is writing history in a way that guarantees his place in it, but as the book is written in textbook-style for laity and does not present new research, whatever Whiggishness we accuse Maxwell of, we can assume that he does believe the kinetic theory of heat to be approximately correct and coherent.

It is in this book that we find the canonical presentation of the demon, which does not appear until three pages from the end of *A Theory of Heat’s* more than three hundred and forty. It is located at the end of the final chapter, which “is devoted to the explanation of various phenomena by means of the hypothesis that bodies consist of molecules, the motion of which constitutes the heat of those bodies” (1871, vi).

This suggests that the demon is not meant as a pillar of justification for the theory, but as a way to help us understand the theory, assuming that it is true. What follows is some textual evidence
for this interpretation. Maxwell says at the beginning of the chapter in which the demon appears: “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, 308). It seems that Maxwell considers the main thesis of the theory of heat to have been demonstrated by this stage of the discourse. He begins asking more interpretive questions regarding the nature of such energy, for example, whether it is potential or kinetic. Maxwell answers that even if it is potential, we must still explain heat transfer by the motion of heat from one location to another. So he concludes that heat energy is kinetic energy insofar as we can investigate it. But kinetic energy of what? To explain this, he turns to the molecular theory of matter. He points out that many authorities in the field have already claimed that the motion of molecules can account for the properties of that which they constitute, where a “molecule” is the smallest part of something that retains the properties of that something (Maxwell 1871, 313). In these molecules, the amount of matter that makes up each molecule is constant for each type of thing. In other words, molecules of gold share a common amount of matter, which is different from the amount of matter common to molecules of water. There are also different degrees of strength possessed by the bonds that keep molecules together. Solids have strong bonds which keep molecules from moving too much in place. Liquids and gases have weaker bonds that allow their molecules to move more freely and intermingle with other types of liquids and gases (although “encounters” between particles last longer than their mere physical contact, and so in liquids there is no time at which molecules are not “encountering” other molecules). All of this, Maxwell claims, has been supported experimentally before the time of writing (313).

Now he considers the situation specifically with respect to gases. Our everyday experience is with the average or conglomerate motion (or energy) of gas, rather than with that of its molecules. Tracing the individual path of a molecule will not be enlightening concerning the motion of the whole, and considering individual molecules might therefore lead us to what appear to be counterintuitive consequences. In his 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” (315-16). Maxwell employs some examples to clarify this distinction between levels of analysis. First he compares the statistical and the deterministic modes of analysis to the same in studying
the effects of education. We may consider the overall impact of education on society, or we may consider the experience of a single student. While one student has an emphatically negative experience, this need not be representative of the overall effect of education on society. In fact there may be no student whatsoever whose experience reflects the average effect of education on his or her society. Maxwell then draws another analogy, this time using the average accuracy of a firing squad versus the path of a single bullet (316). Again, perhaps there is no bullet whose accuracy reflects the general accuracy of the firing squad, yet both are sensible as objects of study. Maxwell is obviously appealing to the imagination of his readers. He wants us to understand what the statistical theory of heat applies to in terms of our everyday experience. He wants us to see that the statistical theory does not compete with everyday experience, as it employs a sort of analysis that cuts across it.

To prove that the molecular approach to heat does not contradict everyday experience, Maxwell shows that the experimental generalities concerning gases of his day can in fact be derived on the basis of the new theory of heat. In this vein, Maxwell demonstrates how Boyle’s law, Guy-Lussac’s law, Charles’s law and Dulong and Petit’s law all follow from the molecular theory (321-333). We emerge from this chapter with a powerful two-way connection between theory and experience. First, our everyday experience provides material from which to create imaginary examples that help us understand how the theory relates to the world as we know it via analogy. These include the examples of education and bullet accuracy. Second, the theory helps to explain empirical regularities (like Boyle’s law, etc.). Even if this dual connection between theory and experience is not a general desideratum of scientific explanations, I think it certainly lends credence to Maxwell’s theory.

The actual text of the thought experiment is situated after the derivation of the above laws and everyday examples that make the statistical components coherent with the deterministic. “Before I conclude,” Maxwell writes, he “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, 338-339).

This is the thought experiment. Following the above-mentioned examples that display the difference between statistical and deterministic properties of the same system, we can see it clearly forms part of the same chapter-long project. And to remind us, 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, 339)

Like the derivations which produced the experimental regularities from the molecular theory, this thought experiment deals with the relation between the kinetic theory of heat and an empirical law; in this case, the second law of thermodynamics. However, a more nuanced approach is necessary here than with the previous examples. This is because Maxwell will not (or cannot) straightforwardly derive the second law from the kinetic theory of heat. Nevertheless Maxwell must provide something, since the theory suggests a result that seems contrary to our
everyday “experience of bodies” (339). Unlike the other theoretical derivations of laws, this connection between theory and empirical regularity proceeds neither by citing empirical evidence nor by mathematical reasoning, but rather by means of example. This is perhaps because the other options are not available: Maxwell does not derive the new interpretation mathematically, and in fact he never presented a mathematical treatment of the second law, at least not in any of his written works or correspondence (Garber, Brush and Everitt 1996, 59). Also, it would be difficult to 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. So instead, Maxwell invents this thought experiment as a tool to assist the imagination. It does not work by simple analogy, as with the average improvement of students or the accuracy of bullets. Considering the accuracy of bullets tells us that statistical regularities can capture the same world that the laws of physics do without contradicting it. The demon seems to do something more. It lets us try on the shoes of the demon and see something for ourselves.

How should we interpret the role of the demon, then? The above contextualization was not meant to lead unequivocally to the one unique role played by the thought experiment, for Maxwell, his contemporaries, or for modern readers. A justificatory role can be played if the demon produces evidence for accepting Maxwell’s theory, or a consequence of the theory. And while the thought experiment does not introduce new concepts, confirm a law, or demonstrate the existence of a regularity, it is easy to see how it could be used as evidence if we portray it as an answer to an objection against the theory. The objection is that the statistical interpretation of thermodynamics makes possible perpetual motion machines, which we know are impossible. Insofar as the demon justifies the theory as a whole from this criticism, it could contribute to scientific knowledge. I want to suggest, then, that this thought experiment performs at least the following two roles: first, it is an example that makes clear some of the empirical content of Maxwell’s theory by relating it to an imaginary experience. Second, it provides some evidence for Maxwell’s theory by virtue of the fact that such sense can be made in the first place.

I will discuss each of these roles, beginning with the second. Häggqvist (2009) argues that a thought experiment justifies new knowledge in the following way. When we perform a thought experiment, we emerge with some new degree of belief with respect to a proposition. Perhaps
this is the result of glancing into Plato’s heaven or manipulating a mental model or executing an argument. However it is done, it seems that sometimes, when the new degree of belief is high enough, we are licensed to say that we emerge from a thought experiment with a new justified belief. And this proposition which we now justifiably believe can be used as a premise in an argument for or against a theory or theoretical claim. This is the case, in fact, for both laboratory and thought experiments, which justify propositions that then become premises in arguments (2009, 62). While there are many ways to fill out this schema, I will only consider the two I find most plausible. They interact with the theory deductively, but in other cases we might see a different sort of relation. Here is the first way to see the thought experiment as playing an evidential role. Consider the following argument:

If Maxwell’s theory is true, then the second law of thermodynamics has merely statistical validity.

The second law of thermodynamics cannot be merely statistically valid because this would permit perpetual motion machines, which are impossible.

Therefore, Maxwell’s theory is not true.

Maxwell uses the demon to deny the second premise by showing that it is indeed possible that the second law of thermodynamics is only statistically valid without allowing for the possibility of perpetual motion machines. By removing this argument against the theory, he has increased its warranted assertability, probability or intuitive acceptability (depending on your epistemology). For example, if we choose to interpret the thought experiment as increasing subjective probability, then assuming that the number of different viable competitor theories is finite, the thought experiment increases the probability of Maxwell’s theory by eliminating competitors (for a formal treatment of this “No Alternatives Argument,” see Dawid, Hartmann and Sprenger 2015). In this way Maxwell’s demon can be seen to provide some weak probabilistic evidence for his theory.

We can also portray the evidential support provided by the thought experiment in a positive way instead of by counterargument. If we can show that the second law of thermodynamics has merely statistical validity, this directly supports Maxwell’s theory, and perhaps the demon does
show this. We can portray the evidential relationship this way: Maxwell contends that matter is molecular, and this is a foundational assumption of the theory that he uses to derive important empirical regularities. The more regularities he derives, the more powerful is this assumption as an explanation. Since Maxwell cannot or will not derive the second law of thermodynamics mathematically, he uses the demon to show how the second law is coherent both with the statistical theory and everyday experience. And this supports the foundational assumption, although weakly. This interpretation of the thought experiment may explain why Maxwell wrote in an undated letter to Tait that the demon shows “that the 2nd Law of Thermodynamics has only a statistical certainty” (Knott 1911, 215).

We could consider other evidential roles, but Maxwell does not raise the demon when he presents the evidence supporting the statistical theory of heat. He raises it afterward, when the theory has already been argued for and applied in many contexts. The thought experiment appears when Maxwell considers how to understand the theory as it applies to experience in general. Because of this, whatever justificatory upshot we can attribute to the thought experiment, I think that in the context of A Theory of Heat, Maxwell’s main purpose was to clarify the consequences of his theory by relating them to imaginary experiences. This role concerns scientific understanding rather than knowledge. I do want to reiterate, however, that the demon could and perhaps was used, simultaneously, as (weak) justification for the kinetic theory of heat, in one of the ways mentioned above. It can be used this way, and surely it has been used this way in the time since Maxwell. But that does not conflict with the idea that there is another complementary use of the demon, which makes possible the evidential use.

I will now consider this other role, which concerns how we understand Maxwell’s theory. I will take up a lot of space examining this role, because it has traditionally been skimmed over by philosophers of science, except for Brown, who uses this role (for Maxwell’s demon, no less) as an attack against Norton:

Norton is right in one sense in this example when he claims that the picturesque details play no role in the argument: the demon is irrelevant to the derivation of the conclusion. But the demon is not irrelevant to the understanding of that conclusion, and that’s the thing Norton’s account misses… I realize that to talk of ‘understanding’ is to wade into
murky waters and to court mystery mongering...The [demon tells] a story in which the events to be explained make sense. We see how things could come about—not how they actually do or did come about. (1992, 274)

Starting now, this dissertation is an attempt to clear up those murky waters.

Here is how we can reconstruct the thought experiment as an attempt to increase scientific understanding. Maxwell invites us to consider what it means to identify heat with the average kinetic energy of particles. This is an abstract theoretical equivalence that is at first difficult to understand. For example, we might ask what it means with respect to our common experience with heat: how does fire make particles move faster? How does a warm liquid cool down? Or we can ask more theoretical questions: can molecules themselves be hot or cold? Is there an upper limit to heat if there is an upper limit to the speed of molecules? Does the kinetic theory of heat unify conduction, convection and radiation as kinds of heat transfer? The more questions that Maxwell’s theory answers, the better his theory is as an explanation.

Explanations are closely related to understanding, although they are not interdefinable (see Chapter 7). Some have argued that if you can explain something, you can render it understandable, and whatever makes something understood can serve as an explanation. As we saw above, Maxwell spends the final chapter of *A Theory of Heat* on questions like the above, and doing so displays the explanatory potential of his theory. One strength of his theory as an explanation is that it enables us to represent a natural phenomenon mathematically. If heat is the average kinetic energy of particles in motion, then heat can be portrayed as something that follows laws which take a statistical mathematical form. These laws can be used to explain some characteristics of heat, such as why fire causes pain (above 44°C particles move fast enough to break apart protein structure, and pain is the response of our nervous system to the resulting cell and tissue damage. See Singer, Taira and Lee 2014, 808). Mathematical representation also allows those who follow Maxwell to predict previously unknown characteristics of heat. This is one reason it is a good explanation. Yet some thermodynamic phenomena are difficult to characterize using mathematics, for instance the second law and the corresponding inability to create a perpetual motion machine (in the sense of a perpetual heat engine). Perpetual motion machines seem fundamentally impossible, not just highly improbable. According to Maxwell’s
theory, our inability to create such machines is not due to a nomological impossibility, but rather a statistical unlikelihood that is so difficult to manifest that we should never expect to experience it. But this sounds like a dodge. What makes something so difficult to manifest that is not itself the consequence of a nomological impossibility? Our imaginations need assistance, and Maxwell’s thought experiment provides it.

We are asked to envisage a tiny being that can see and interact with individual molecules directly. This is something we could never do. To such a creature, “the average state of molecular motion” is a counterintuitive abstraction, since it sees only individual molecules moving at different speeds. In our imagination, 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.” Why not? Because if we were that being, then there would be no reason we could not. It is difficult to imagine in detail 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. 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, although a natural process analogous to the demon’s intentional behaviour, with actual molecules, is very hard to imagine. As far as we know, there is nothing that operates like such a being.

Hopefully by this point, the merely statistical nature of the second law of thermodynamics makes sense to us. That is, we understand how the law could be violated: anything that causes molecules with high energy to make their way into the same part of a closed region will do. And we also understand why it would be so difficult to find something in nature that recreates what the demon does artificially. There is nothing of that size which can control frictionless shutters in thermodynamically closed systems, and perhaps there could not be. Furthermore, all known natural but non-intentional systems that can separate fast moving from slow moving particles (like semi-permeable membranes) slow down fast moving particles and generate heat instead of decreasing entropy. 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 practically negligible. This recovers our everyday
experience with heat engines, and also clarifies the second law by interpreting it according to Maxwell’s kinetic theory of heat.

Maxwell could have simply proclaimed that the new theory makes the second law of thermodynamics statistically violable. But instead he wanted his readers to understand what this meant, and how this was consistent with our experience of heat engines:

Maxwell's main point was that if the shutter were left open for a long enough time interval…it is possible, though not highly probable, that similar temperature or pressure differences could be established by statistical fluctuations [alone]. Of course the time interval needed would be impractically large and we could not depend on such statistical violations of the second law of thermodynamics to get practical work tasks done. (Leff and Rex 2002, 408-409)

The demon does not require us to visualize statistical likelihoods compounding over immense timeframes, and this makes it attractive as a tool for increasing understanding. It is certainly possible that Maxwell first devised the thought experiment for his own understanding of how the statistical theory related to the world via the second law, and then tried it out on Tait and Strutt before including it in his textbook. And this would be consistent with the chronology presented at the beginning of this section, which seemed to indicate a similar search for understanding. If the thought experiment helped Maxwell understand the relation, perhaps it would do the same for others as well.

Further support for my reading of Maxwell’s demon comes from Garber, Brush and Everitt, who write, “In view of the great popularity of his demon, it may seem surprising that Maxwell did not develop a quantitative treatment of irreversibility based on this idea. Unlike Boltzmann, he did not try to formally demonstrate his statistical interpretation of the second law, rather he explored its general physical and even theological implications” (1996, 59). If the thought experiment is primarily an attempt to increase understanding, then it makes sense that Maxwell did not explore the second law mathematically, but rather chose to proceed by imaginary example.

Finally, the demon has not been accepted by all of Maxwell’s contemporaries, or those who followed. Instead, it spawned a series of critical responses “that has no apparent end” (Earman
and Norton 1998, 435). This, however, is not because the demon failed to illustrate Maxwell’s theory, or because critics believed the second law was sacrosanct (although for some it was). It is because the very idea of the demon raised new questions, for example, concerning what other properties the demon must have if it were to exist. This is important question because if we could make an actual device with those properties, we could take mechanical advantage of the statistical nature of heat. Note, however, that this is an entirely separate issue, one that does not conflict with the idea that Maxwell’s demon is a way of coming to terms with the statistical nature of the second law of thermodynamics, either for Maxwell, his peers, or today’s students.

In conclusion, Maxwell’s thought experiment is capable of playing at least two roles. One is related to the production of knowledge. Initially we considered how the demon might play a justificatory role, and I concluded that there are several argument types that it could fit into, each of which lends support to the theory. And in each of those argument types the demon makes plausible a proposition such as: “the second law of thermodynamics is statistical only, yet will not lead to perpetual motion machines.” Nevertheless this role for the demon, successful or not, cannot be its only role, as it depends on another that is related to the production of understanding. The thought experiment helps us understand the second law of thermodynamics as merely statistical, and only once this is achieved can we use the demon in arguments for the theory, and against counterarguments to the theory.

We are left wondering what scientific understanding is. I will continue to explore case studies that exemplify scientific understanding, and common features will emerge as we go. The full account can be found in Chapter 7, section 5. For now, I will continue to characterize understanding as an epistemologically desirable state that is not propositional knowledge. When we come to understand something, it becomes intelligible or meaningful for us, and we can use it to achieve our ends. It is less a case of learning something new, than gaining a new ability.

2 The Clock in the Box

The clock in the box is a thought experiment of Einstein’s, presented at the 1930 Solvay conference, which tries to show the incoherence of quantum theory by targeting the uncertainty principle proposed by Werner Heisenberg in 1927 (see Kragh 2002, 212-213). Einstein
apparently presented the clock in the box “to evade the fourth Heisenberg uncertainty relation” (Treder 1975, 135), that is,

\[ \Delta E \Delta t \geq \hbar. \]

In other words, the product of the “uncertainty in the knowledge” (Heisenberg 1930, 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 (Plank’s constant divided by 2π).

Bohr is said to have been preoccupied with this thought experiment until the day he died (Kalckar 1996, 286). This is surprising because according to Bohr’s account of the exchange, Bohr successfully refuted Einstein (Bohr 1949, 226). At the same time, Bohr’s life-long engagement with the clock in the box makes sense if he appreciated the widely recognized fact that his rebuttal was only convincing enough.

The clock in the box was not technologically realizable when Einstein and Bohr developed the thought experiment, and modern versions have not been realized (see Dieks and Lam 2008, 838). Here is Einstein’s thought experiment as presented by 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. 7). 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). Subtract the two, and we get the energy of the photon. Since we know the exact time it was emitted, we obtain a contradiction with (1).

The thought experiment was “quite a shock to Bohr,” nevertheless with the “next morning came Bohr's triumph” (Rosenfeld, quoted in Pais 1982, 446-447). Different authors present Bohr's triumph in different ways, but it is agreed that his first step was to provide a detailed analysis of
the experimental set-up. To that effect, he modified Einstein's scenario. A “pseudo-realistic”
drawing (Bohr 1949, 226) helped Bohr to demarcate the crucial differences between Einstein's
set up (Fig. 7) and his own (Fig. 8).

![Fig. 7: Einstein’s Clock in the Box, Depicted by Bohr (Bohr 1949, 225)](image1)

![Fig. 8: Bohr's Pseudo-Realistic Drawing of the Clock in the Box (Bohr 1949, 227)](image2)
What Bohr did was look closely at the weighing process required to determine the mass of the emitted photon. This might be surprising, because that particular process seems like the least problematic aspect of Einstein’s thought experiment. Here is a reconstruction of Bohr’s mathematical derivation of the uncertainty principle from the thought experiment (following Bishop 1999), which relies heavily on the above diagram and some considerations concerning measurement.

Given the experimental set-up as shown in Fig. 8, and assuming the Heisenberg relation for position \( q \) and momentum \( p \),

\[
\Delta p \Delta q \geq \hbar,
\]

Bohr argued as follows. The more precise our knowledge is of the level of the pointer that represents the position of the box in accordance with the box’s weight and momentum, the less precise is our knowledge of the momentum of the box, and vice versa. When weight is added to the bottom of the box, there will be a change in its momentum. This change in momentum—symbolized as \( I \) for impulse—over the time it takes to bring the pointer back to the initial level by adding weights of different magnitudes, is given by

\[
I = T g \Delta m
\]

where \( T \) is the time of the weighing process, and \( g \) is the gravitational constant. Bohr then makes what he thinks is a harmless assumption, namely that the uncertainty in the box’s momentum (\( \Delta p \)) must be less than the total change of momentum (impulse) impelled over the entire measurement procedure. Expressed symbolically, we have

\[
\Delta p \approx \hbar / \Delta q < I = T g \Delta m.
\]

Bohr goes on to note that weighing the box will necessarily involve the presence of a gravitational field. 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. Specifically, there will be a change in time, \( t \), for every change in position within the gravitational field according to the following equation

\[
\Delta t = T g \Delta q / c^2.
\]
Isolating for $T$ in (6) we get

$$T = \frac{(\Delta t \, c^2)}{(g \, \Delta q)}.$$  

If we introduce this into (5), we obtain

$$\frac{\hbar}{\Delta q} < \frac{(\Delta t \, c^2 \, g \, \Delta m)}{(g \, \Delta q)},$$

which reduces to

$$\hbar < \Delta t \, \Delta m \, c^2.$$  

And given (2), Bohr concludes

$$\hbar < \Delta t \, \Delta E$$

which confirms (1). Presented with this reply, Einstein permanently “accepted fully all the Heisenberg relations” (de la Torre et al. 2000, 54), and never brought up the matter again (he said afterwards that the correctness of the uncertainty relation was, from his point of view, “rightfully regarded as finally demonstrated”; Einstein 1949b, 663).

This is surprising because neither version of the thought experiment is especially convincing. A quick look at some of the more or less recent treatments of the clock in the box makes this sufficiently clear, and also demonstrates the theoretical fertility of the thought experiment.

Treder (1975) argues that there is no relation between Einstein's red shift and the Heisenberg relation for energy and time, which is assumed by Bohr in his reply. Consequently, Treder (1975) presents a reformulation of the clock in the box. Instead of weighing the system in a gravitational field, the equation

$$Q = Ne$$

is employed, for the total charge $Q$ with $N$ particles with identical specific charges $e/m$, according to

$$c^2(E_1 - E_2) = \frac{\delta E}{c^2} = \delta Nm = \frac{(m/e)}{\delta Q}, M = Nm.$$
If we add for the total charge $Q$ of the box, given by

$$Q = Ne = (e/m) \sum m = (e/m) m M = (e/m) m (E/c^2), \tag{13}$$

then we can determine precisely the energy of the system.

Alternatively, Unruh and Opat (1979) derive the relativistic slowing down of the clock by means of Hamiltonian mechanics in order to demonstrate that Bohr was right in his disagreement with Einstein, although wrong in the way he justified it. Within the context of Hamiltonian mechanics, the slowing down of the clock follows “directly from, and is a necessary consequence of the possibility that energy has weight” (Unruh and Opat 1979, 743).

De la Torre et al. (2000) argue in another direction, namely that Bohr was equivocal about the meaning of $\Delta T$ in his presentation of the thought experiment. Hnizdo (2002, L11) disagrees, “Bohr’s meaning of $\Delta T$ is the same as that assigned to it by Einstein, contrary to what [de la Torre et al.] assert.” This debate is still open.

Hilgevoord (1998) reminds us that the uncertainties of position, momentum and energy can be represented as spreads of the probability distributions $|<q|\psi>|^2$, $|<E|\psi>|^2$ and $|<p|\psi>|^2$, respectively. But “an uncertainty in time cannot be construed this way” (Hilgevoord 1998, 397). And the phenomenon of Bohr’s thought experiment implies just such a move in his reply to Einstein. According to Kudaka and Matsumoto (1999), this implication amounts to a big problem in Bohr's response to Einstein, because “Bohr and Rosenfeld stressed the principle that every proper theory should provide in and by itself its own means for defining the quantities with which it deals,” and in the case of the clock in the box, “we require the concept of a clock, and the clock cannot be independent of the various physical laws” (Kudaka and Matsumoto 1999, 1244).

Hilgevoord (1998) can be read as an attempt to meet this requirement on Bohr’s behalf. In order to give meaning to the uncertainty in time, Hilgevoord re-describes the clock as a moving position (the end of the hand of the clock) that varies with an angle $\varphi$ over the interval $[0, 2\pi]$. What makes this a clock is a strict determination of the momentum of the pointing hand, which implies a probability spread if we try to measure the angle $\varphi$. It is concluded that “there is an uncertainty relation between the accuracy with which a physical system can indicate time and the
width of the energy distribution of its quantum state…This uncertainty relation does not, in itself, rule out the existence of an ideal clock; it says only that the energy spread of an ideal clock must be infinite” (Hilgevoord 1998, 399-400).

The proper way to derive the fourth uncertainty relation is then to see that if “the total momentum is certain, all position variables are completely uncertain” (Hilgevoord 1998, 400). Then, if we consider the shutter inside the box “as the hand of a clock passing over the hole” and we weigh the system in the beginning to determine the exact total energy of the system, we see that the time has necessarily become “unsharp” and the time of emission of the photon cannot be known precisely (Hilgevoord 1998, 402).

Treder reminds us of yet another problem. According to Bohr and Heisenberg, all that (1) says is,

\[ (1') \quad \Delta T > \frac{\hbar}{\Delta E}. \]

That is, you cannot know \( \Delta E \), that is, \((E_1 - E_2)\), in less than \( \Delta T \) without increasing the uncertainty of \( \Delta E \). Therefore, the fourth uncertainty principle only applies to dynamic (non-static) systems (Treder 1970, 21-2). At this point the question can be raised as to how we might conceive the clock in the box thought experiment of Einstein’s as dynamic. After the photon leaves the box, there will be a time of oscillation that is damped by the spring. Some of the box-spring system’s energy will be lost as heat in this process. We are asked therefore to imagine not just one system, but many clocks-within-boxes, each of which ejects a photon at the same time. The oscillation periods will be different for each clock in the box, because of the varying rates of heat loss/transfer/damping of the spring according to statistical thermodynamics. Therefore, there will also be different final energies of the systems. The shorter the vibration state time, the bigger \( \Delta q \), and the proportion will vary according to the momentum-position uncertainty relation. In other words: the shorter the time needed to reach the steady state, the bigger \( \Delta E \) will be. And we will need at least the interval \( \frac{\hbar}{\Delta E} \) to separate the measurements of \( E_1 \) and \( E_2 \) at \( T_1 \) and \( T_2 \). Thus, Treder concludes that the “fulfilment of the fourth uncertainty relation is therefore a consequence of the quantum properties of the spring” (Treder 1970, 24).

It seems fair to conclude that “regardless of whether the box is moving freely or undergoes a harmonic motion, and regardless of whether we base our predictions on the determination of \( q \) or
we find (1) confirmed (Dieks and Lam 2008, 841). However the manner in which the uncertainty principle is best demonstrated from the clock in the box is an open matter.

Rather than focus only on the thought experiments of Bohr and Einstein, I want to look at the whole episode. What is it about? As with Maxwell’s demon, there are several ways to characterize this chain of thought experiments. And again, two roles I will focus on are the attempt to ground new knowledge and the attempt to increase understanding.

According to Bishop (1999) the same thought experiment was performed by Einstein, then repeated and improved by Bohr. It has different conclusions in each case, and therefore if Norton is right about thought experiments, we have one thought experiment being equivalent to two arguments, since two identical arguments cannot have different conclusions. If Norton tries to escape by representing Einstein and Bohr’s thought experiments as different arguments, then there is no reason to think that Bohr addressed Einstein’s thought experiment at all. He merely changed the topic, and introduced a similar imaginary scenario that does obey the uncertainty principle. In this case, Einstein and the physics community were epistemically irresponsible for listening to Bohr (Bishop 1999, 540). According to Bishop, therefore, Bohr’s response only makes sense construed as a repetition and improvement of the same thought experiment that Einstein introduced.

This disagreement between Bishop and Norton portrays the thought experiment as an attempt to disprove or vindicate the uncertainty principle. In other words, the debate assumes that it is a thought experiment concerned with the epistemic status of that principle. Let us assume that this is a legitimate portrayal in order to consider the role of the thought experiment as evidential. Several ways this role can be played are made clear by considering the respective positions of Norton and Bishop.

If Norton is right, this thought experiment is an argument made by Einstein to show or derive a contradiction with the uncertainty principle. Bohr sees this and answers by performing a different (and better) argument which disproves the first by using some hidden background assumptions of Einstein, who failed to consider the effects of the gravitational field in measuring the weight of the box. Bohr’s new argument shows that if we really performed an experiment like the one Einstein suggests, we would recover the uncertainty principle exactly.
Norton’s response is reasonable. Bishop seems to be denying that there are arguments (and experiments) that are not the same and nevertheless disprove one another. According to Norton, Einstein puts forward an argument, and Bohr presents a new argument that both demonstrates the invalidity of Einstein’s and establishes the uncertainty principle. Bohr’s argument deals with a similar imaginary scenario that concerns the same physical principle, but that is more or less where the overlap ends. We can even reconstruct the exchange logically: Einstein uses the thought experiment to justify a modal premise, namely that the clock in the box scenario is possible as he describes it. If this is successful then the uncertainty principle is violable. Bohr denies Einstein’s conclusion by denying that the scenario in the premise (including its implicit, dynamic measurement procedures) is actually possible. In other words, he denies the truth of the modal premise.

Nevertheless, Bishop might be correct that the thought experiment is meant to disprove or validate the uncertainty principle while not being an argument. What then, accounts for its success? Bishop “agrees with Sorensen that there are many similarities between thought experiments and real ones” (1999, 535), so perhaps Bishop would agree with the conclusion of Chapter 3 that we should focus more on the experimental features of the case rather than the logical relations between its assumptions and conclusion. We vary the properties of a scenario in our minds to see how this affects some other properties we believe to obtain in that scenario, using tacit knowledge and perhaps rational intuition. Performing the thought experiment provides some mental presentation that can be used as evidence for or against the uncertainty principle. Einstein’s thought experiment presents a scenario that we might judge as plausible. And the plausibility of that scenario casts doubt on the uncertainty principle. Then Bohr’s thought experiment presents a different scenario that we accept as plausible and which casts doubt on our previous conclusion.

Brown has worked out a detailed account of this sort of interaction with respect to theory confirmation. To see how it works, grant that the variations of the thought experiment are meant to justify scientific conclusions and thereby lead to new knowledge. To do this we characterize the responses as “counter thought experiments” (Brown 2007a). This kind of thought experiment will “challenge neither the premises nor the concluding inference. Instead, counter thought experiments deny the phenomena of the initial thought experiment” (Brown 2007a, 157). The
“phenomenon” referred to here is the mental presentation that results from the execution of the proposed manipulations of the imagined scenario. (For more on Brown’s view of “phenomena,” see Brown 1993b).

In clarifying the concept of a counter thought experiment, Brown uses the following visual aid to highlight the difference between a narrow and broad conception of experiment (Brown 2007a, 158):

Theory and Background → Phenomenon → Result

An experiment (thought or otherwise) considered broadly contains the whole progression, where the arrows are meant to represent the general idea of a progression through time. A narrow conception of the same experiment contains only the middle section: the production of the phenomenon. A counter thought experiment does not concern itself with background suppositions or whether the conclusion of a thought experiment follows validly from its premises. Rather, a counter thought experiment simply denies that the phenomenon in a thought experiment would occur as claimed. The clock in the box appears to be such a counter thought experiment.

For Bohr, the key to Einstein’s challenge is the fact that weighing is measurement, and measurement requires some action on the part of measuring agents. In the quantum realm this action and its influence on the measured system are important because they are inseparable from the system, and therefore inseparable from the phenomenon of the experiment. He writes: “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).

Bohr aims at the weighing mechanism in order to remind us of the difficulties specific to measuring things on the quantum scale. Thus Bohr tries to render the phenomenon of Einstein’s thought experiment counterintuitive. In his choice of modifications in the experimental set-up, he targets the possibility of measuring with arbitrary precision the energy of a photon at a precise time. This is the crucial part of the experimental phenomenon for Einstein. In the second step, Bohr uses his modified version of the clock in the box to derive the Heisenberg relation for
energy and time. As we have seen, his derivation may be problematic. Yet his counter thought experiment is nevertheless successful in having caused the relevant parties to doubt the phenomenon of Einstein’s experiment.

To be a successful counter thought experiment, for Brown, is to deny convincingly the plausibility of the target thought experiment’s phenomenon (2007a, 170). Brown helpfully provides a list of questions to determine the plausibility of the phenomenon of a thought experiment (and its counter):

1. How reliable is the initial thought experiment in the narrow sense (that is, do we think the phenomenon could obtain, without considering our background theories in much detail)?

2. How strong is the assumed background to the thought experiment?

3. How similar is the phenomenon of the thought experiment to things we know and trust?

4. How plausible is the phenomenon of the counter thought experiment?
   a. How absolutely plausible?
   b. How relatively plausible?

“Absolute plausibility” refers to how plausible the phenomenon would be outside of the context of the counter thought experiment. “Relative plausibility” refers to how plausible the phenomenon is compared to the phenomenon in the initial experiment: is it more or less plausible? Let us ask these questions of the clock in the box:

1. Einstein’s thought experiment is reliable in the narrow sense; it has very few parts (a clock, a box, a shutter and a light source) whose combination we unreflectively think is unproblematic. We know how clocks work and we know how to weigh boxes. For this reason it is easy to understand, and seems plausible.
2. The background to this thought experiment is very minimal. It is little more than the uncertainty principle itself and $E = mc^2$.

3. The phenomenon is very similar to things we already understand; we can use spectrometers to measure energy, we know that mass interacts with gravity to make up the quantity we call weight, we understand how clocks work, and we have seen light escape through openings.

4. How plausible is the phenomenon of the counter thought experiment?
   a. It is not very plausible in the absolute sense. Although his illustration is helpful, it is difficult to conceive the situation Bohr describes in detail. It requires changes of momentum in gravitational fields and also the application of weights with hooks that weigh so little they might not even be visible.
   b. It is somewhat plausible in the relative sense. Compared to Einstein's experiment, Bohr's is complicated and there are several places where mental uncertainty can creep in, as we saw with the many reactions to the thought experiment detailed above. However it feels more solid than Einstein's for all its added detail, and especially its inclusion of relativistic effects in a gravitational field.

Given 4.a., we should wonder why Bohr's counter thought experiment could have been effective for the scientific community. Brown has an answer: a counter thought experiment fails only when

\[
\text{Prob (counter phenomenon)} / \text{Prob (initial phenomenon)} \ll 1
\]

In other words, a counter thought experiment will only fail when the relative probability of the counter thought experimental phenomenon is significantly less than the probability of the initial thought experimental phenomenon. This is Brown's "ratio test." Brown uses probability here in the sense of signifying the degrees of belief of an agent or community, and he is happy to "admit and even stress that this is not intended to be realistic" (Brown 2007a, 174).
It is important to note that “significantly less” is not as vague as it seems: compare the situation of the audience at the Solvay conference to a jury attending a trial. (Treating scientific audiences like juries goes back at least to Robert Boyle and the Royal Society; see Sargent 1995). A counter thought experiment needs only to inspire reasonable doubt in the phenomenon of its target (Brown 2007a, 174). If a counter thought experiment can inspire reasonable doubt about the phenomenon of a thought experiment, we will lose our enthusiasm for its conclusion. The flimsier a thought experiment’s evidence, the easier it is to establish reasonable doubt.

Now, is Bohr’s thought experiment plausible enough to cast sufficient doubt on Einstein’s? History provides the short answer: Yes. With respect to the task of demonstrating reasonable doubt, he was successful.

What about the modern versions of the thought experiment? If Einstein’s thought experiment had some relatively high degree of probability for the scientific community, and Bohr’s had less absolute probably but enough relative probability to inspire reasonable doubt, this raises a problem for the ratio test with respect to the modern performances of the thought experiment. If the absolute probability of Bohr’s experiment can be significantly less than Einstein’s and still succeed, this implies that a thought experiment which denies the phenomenon of Bohr’s can be even less (absolutely) probable, while still being effective. With only a few iterations we see that the longer a chain of counter thought experiments stretches, the less absolutely probable our thought experiments are permitted to become, without any loss in effectiveness. We could predict that over time we would see wackier and wackier thought experiments. But instead we see the opposite; modern instantiations of the clock in the box have become more and more (absolutely) plausible. This does not disprove the ratio test as a success-criterion, but it does show that there is something else going on here.

Here is another proposal to make sense of the episode. Each counter thought experiment in the chain is an attempt to connect the uncertainty principle either to experience or other parts of physical theory. Some of the experiments cast doubt on the previous ones, but they do not do so by comparative plausibility. Each link in the chain establishes a connection. While connecting things is easy, making a useful and satisfying connection is not always so easy. A counter thought experiment is a good one when it connects the right things in the right way. It succeeds
over its predecessor because it connects the same things in a better way, or connects better things. Which connections are most appealing will be relative to the individual, although in science, tastes will often align because values often align. Einstein connected the variables that appear in Heisenberg’s uncertainty principle to experience in a way that violates the proposed relation between them. But his imaginary scenario was not detailed enough. The connection that he proposed could not in fact be made. This is something that all of the counter thought experiments since Einstein’s original thought experiment have agreed on. Bohr’s thought experiment has been rejected by some, but not his conclusion, and it is the conclusion that matters for the scientific community. This is not new in the history of science.

Still, we want to make sure that the world really is the way that the uncertainty principle says that it is. The fact that the clock in the box is only imagined does not matter in this case because the thought experiment concerns a point of possibility. We want to know if it is impossible to violate Heisenberg’s relation in principle. For this reason the clock in the box is still discussed.

What would scientists achieve if they agreed about how exactly to show that the clock in the box cannot violate the uncertainty principle? Several things: first, such an accomplishment would serve as evidence that the uncertainty principle is true by using the completed thought experiment to defeat Einstein’s sort of challenge decisively once and for all by denying his premise (á la Norton) or the phenomenon of his thought experiment (á la Brown). On the other hand, the uncertainty principle is already scientifically well-established, and was, even in 1930. Perhaps there is some other role the clock in the box is playing. And I think there is. This thought experiment is useful for increasing the empirical content of the uncertainty principle. The principle simply sets a maximum level of precision for measurements of pairs of conjugate variables, or knowledge of those quantities, or the definition of those variables, or the “statistical spread” of those variables (Hilgevoord 2006). But the principle does not say why that limitation exists. So how do we determine the empirical content of the uncertainty principle?

As we saw above, Einstein proposed an imaginary case that seemed not to obey the Copenhagen interpretation of Heisenberg’s uncertainty principle. About the exchange Bohr says, “there could be no other way to deem a logically consistent mathematical formalism as inadequate than by demonstrating the departure of its consequences from experience or by proving that its
predictions did not exhaust the possibilities of observation, and Einstein’s argumentation could be directed to neither of these ends” (1949, 229-230). That is, the only thing that a thought experiment directed against a theory could hope to achieve, according to Bohr’s version of positivism, would be to show that there are possible or actual experiences that contradict the predictions of the theory. Quantum mechanics faces no trouble from the clock in the box, according to Bohr, because an actual performance of the experiment would be in line with the theory’s empirical predictions. The thought experiment is therefore at least partially about understanding Heisenberg’s uncertainty principle; finding whether it applies to the world as it claims to, and determining how it should be interpreted.

However, perhaps the clock in the box was intended as more than a counterexample to the uncertainty principle. For Einstein, the clock in the box was an extension of his long-time attitude to quantum mechanics: “Einstein never ceased to ponder the meaning of the quantum theory” (Pais 1982, 8). Norton writes: “The task of responding to Einstein was taken up by Niels Bohr. The debate in which they engaged was surely one of the monumental debates of the 20th century. Here were two titans of modern physics with quite opposed positions, struggling to establish their view of the meaning of the quantum” (unpublished (b), chapter 29, my emphasis).

I interpret the clash between Einstein and Bohr as one concerning what the uncertainty principle means. Establishing coherent semantic content for quantum mechanics could certainly be characterized as an increase in knowledge, but it is also important to see it as increasing understanding. We would not necessarily learn any new facts about the world, or justify any theoretical claims that were not already justified. Rather we would come to understand what our existing knowledge meant for experience, and get clear on how it relates to other theories.

To summarize this discussion of the clock in the box, we have here a series of thought experiments that can be characterized as performing the same two roles as Maxwell’s demon. On the one hand the thought experiments can be portrayed as trying to provide new evidence for knowledge claims surrounding the uncertainty principle. On the other hand they can be seen as attempting to increase our understanding of this principle. I contend that if these thought experiments play the first role, it is likely because they also play the second. This is because
knowing what a theoretical structure means implies knowing what kind of world it describes, which is crucial for all discussions concerning the accuracy of that structure.

Finally, note that the chain of thought experiments considered as a whole appears to succeed in increasing understanding, whether or not it succeeds in increasing knowledge. Merely by going through the presentation of the various thought experiments, we have gained a better grasp of the empirical content of the uncertainty principle (how it relates to other principles, how it could manifest in possible experiences, etc.), and this is true regardless of whether we have gained any new justification for the principle.

3 Einstein’s Elevator

Einstein reveals the Special theory of Relativity in 1905, and from 1907 until 1915 he works hard trying to generalize it to gravity. Fundamental to this shift is the principle of equivalence, which says that the laws of physics will function equally in frames of reference subject to gravitational forces as they do in those subject to uniform acceleration. This is first suggested in 1907 (Sauer 2004). The field equations are given in Einstein’s 1915 “Die Feldgleichungen der Gravitation.” The most important article for General Relativity is Einstein’s 1916 review article “Die Grundlage der allgemeinen Relativitätstheorie,” which presents the more or less complete theory. In this article, Einstein gives four different arguments for the equivalence principle. One main motivation for accepting the principle concerns the epistemological issue of verification. Since we cannot even in principle tell between the two types of frame of reference (that is, all the physical regularities we experience in one type of frame will also be experienced identically in the other), we can treat them using the same mathematical-theoretical formulas. This is because those formulas describe the regularities we experience, and those regularities will be the same in both types of frame. The elevator thought experiment is our primary way of understanding the equivalence principle in textbooks and educational websites (for example, Wolfson and Pasachoff 1999, 1037; Feynman, Leighton and Sands 1964, 42-45; Wikipedia’s entry on the Equivalence Principle: http://www.wikipedia.org/wiki/Equivalence_principle; and Einstein-Online: http://www.einstein-online.info/spotlights/equivalence_principle). And yet, the thought experiment does not appear as one of the four arguments that Einstein uses in 1916 to establish the principle when it is first introduced.
This is curious because Einstein directly refers to *Galilean* relativity in his 1916 paper, which is introduced and justified by Galileo (1632) using the following thought experiment. We imagine that we are in a ship that is either moving with perfectly uniform speed, or at rest, and there is no way to look outside. Galileo notes that our inability to tell whether we are moving or not from within, either by observation or experiment, demonstrates a type of empirical equivalence between frames of reference moving with uniform speed and those at rest. I presume that Einstein knew of this thought experiment because almost everyone who learns about Galilean relativity (which Einstein would have) encounters it, and also because his elevator thought experiment seems obviously inspired by it (because as we will see below, it mimics it in almost every detail). So why not include it in the 1916 article in *Annalen der Physik*? My conjecture is that this article attempts to provide theoretical-mathematical arguments for the equivalence principle in a way that appeals to those with all the necessary physical and mathematical background, and imaginative strength. Only afterwards, when Einstein is concerned with making the principle of equivalence understood, does the thought experiment appear.

The elevator thought experiment is first published in a popular book written by Einstein in 1916, which is translated into English in 1920 as *Relativity: The Special and the General Theory*. It reappears in the 1938 book *The Evolution of Physics*. The preface to the former opens with these words:

> The present book is intended, as far as possible, to give an exact insight into the theory of Relativity to those readers who, from a general scientific and philosophical point of view, are interested in the theory, but who are not conversant with the mathematical apparatus of theoretical physics…The author has spared himself no pains in his endeavour to present the main ideas in the simplest and most intelligible form.

The fact that the elevator appears in this book and not in the relevant journal articles supports the idea that the elevator is not meant primarily as an argument to establish the equivalence principle, but to provide more general understanding with respect to something that is otherwise established. It allows the reader to understand for herself the principle, and what it means for experience and for physical theory. That is, the elevator is epistemically efficacious, but not because it provides new knowledge. My analysis will follow the same trajectory as with
Maxwell: we have here a scientist who develops something revolutionary in articles, and then publishes a book which attempts to make these results broadly appreciable. In both *A Theory of Heat* and *Relativity*, gaps in scientific knowledge are highlighted. The newly developed theory is presented, including how it addresses those gaps. And finally, the theory and its empirical consequences are considered in a way that makes their import clear to the reader. And it is in these final stages that we find both Maxwell’s demon and Einstein’s elevator.

However if the reader is inclined to see this thought experiment as concerned with knowledge acquisition, he or she is free to do so. One way is to treat the thought experiment as establishing a premise in an argument for General Relativity. As with Maxwell, we can imagine an interlocutor who denies Einstein’s premise that, in reality, gravitational and inertial frames of reference are empirically equivalent. And then we can characterize Einstein as providing an imaginary case that justifies our belief in the following proposition: “it is possible that gravitational and inertial frames of reference are empirically indistinguishable.” In other words, someone could use the thought experiment as weak modal evidence for Einstein’s theory by refuting a counterargument.

And again as above, the thought experiment could be used positively by making the empirical equivalence between gravitational and inertial frames of reference plausible, which would lend some (weak) support to the idea that these frames can be treated mathematically as identical, and this lends support to the theory as a whole. This is how Norton reconstructs the thought experiment. Compared to Einstein’s other arguments for the principle of equivalence, portraying the thought experiment this way makes it look toothless, and accordingly Norton concludes that it is “quite problematic” (1991, 138). While I accept that the thought experiment might have played this role for Einstein, and can still play it for us, I think there is a better way to understand the elevator given its important role in the history and pedagogy of science.

Namely, the thought experiment increases understanding. Further, any evidential role it plays is made possible by this more fundamental role. To portray the elevator as increasing understanding, I cannot reconstruct Einstein’s thinking along the same lines as Maxwell’s, because while it is true that the thought experiment was published after the technical papers that contain the theoretical framework of General Relativity, Einstein first conceived this thought
experiment, which he called “the happiest thought of my life,” in 1907, eight years before he introduced the theory (see Norton, unpublished (a)). One might therefore argue that this thought experiment was the original spark that justified General Relativity, and not an afterthought used by Einstein to explain it (as might have been the case with Maxwell). If this is true, his thought experiment is part of the original justification for the theory, even though Einstein reveals it only later. This makes the thought experiment appear justificatory first and explanatory second.

I do not deny that this “most beautiful thought” might deserve some or much of the credit for General Relativity. But even if the thought experiment was the spark, Einstein still laboured through years of painstaking mathematical effort to justify the principle of equivalence before presenting it. To Einstein then, the thought experiment was not enough to justify the principle to the standards of justification expected by his peers.

I argue that the thought experiment could have increased Einstein’s understanding of the principle of equivalence, even in 1907, long before the theory was justified to his satisfaction. This is because understanding can precede knowledge. Suppose that understanding $x$ is to be able to work with $x$, or to have an intuitive grasp of $x$, and knowing $x$ is something like having a justified true belief that $x$. In this case, there is certainly a sense in which Einstein may have understood General Relativity before he had a true justified belief in it. There is some evidence for this. Consider Einstein’s famous statement:

I believe in intuition and inspiration. … At times I feel certain I am right while not knowing the reason. When the eclipse of 1919 confirmed my intuition, I was not in the least surprised. In fact I would have been astonished had it turned out otherwise. Imagination is more important than knowledge. (1931, 97)

It seems that Einstein may have had some intuitions about the theory, which came as a result of his imagination. Having these intuitions was enough to make him feel as though he understood the nature of gravity, and made him feel confident that he was right. But the fact is he did not make his theory known in 1907 when he had his happiest thought, even though he felt sure it would work out to be correct. He did not yet have the evidence the scientific community would rightly demand. Only when the justification was finally worked out could he finally present it as a candidate for scientific knowledge.
Before we move on to address the understanding that the thought experiment can provide, I should note another role attributed to Einstein’s elevator. It is often characterized as a direct argument for a consequence of the principle of equivalence, which is that light will bend in a gravitational field. This reading, however, is historically problematic. In *Relativity: The Special and General Theory*, Einstein spends an entire chapter on the elevator (chapter 20), and does not mention the extension of the thought experiment to the behaviour of light at all. When this consequence of the principle does appear, it is in chapter 22 (1920, 69-70) and the elevator is nowhere to be found. The use of the elevator as an argument for the bending of light in a gravitational field is therefore anachronistic. Of course, this does not mean that Einstein would have disapproved of extending the elevator in this way; it is a very useful extension. But even if he had, this would not cast any doubt on the dual use of the thought experiment as both a tool for knowledge and understanding, because such an extension would also increase understanding, although about the behaviour of light, and not the nature of gravity.

Now, what type of understanding does the elevator provide, and how? Norton reminds us of Einstein’s situation immediately after the publication of his work on Special Relativity:

> It is easy to forget just how precarious were Einstein’s early steps toward his general theory of relativity. These steps were not based on novel experimental results. Indeed, the empirical result Einstein deemed decisive—the equality of inertial and gravitational mass—was known in some preliminary form as far back as Galileo. Again, there were no compelling theoretical grounds for striking out along the path Einstein took. (Norton 1993, 3)

After years of mathematical toil, Einstein comes forward in 1916 with a paper that introduces the equivalence principle. But presentation is not always explanation, and neither the principle nor the associated mathematics make it immediately clear how planets or people or puppies will act if it is true. It *might* do this, but only if your knowledge of physics, mathematics, *and the strength of your imagination* are such that it is obvious how it can be applied. This was not the case for most of the physics community, save Eddington and perhaps some members of the $\nabla^2 V$ club (Sponsel 2002, 457-458). In performing Einstein’s elevator thought experiment, the imagination is given training wheels. By imagining a scientist (but really, yourself) in a windowless crate
being pulled upward through empty space so that it constantly accelerates at $9.8\text{m/s}^2$, Einstein presents a situation that makes it clear how the equivalence principle would be realized in the world. Objects carried with you into such a situation would act exactly as they would if that same crate were motionless on the surface of the earth. We must imagine ourselves in both scenarios, and compare what we would experience in each using our imagination to see if there would be any difference. If we conclude that there would be no difference in our experience of the two scenarios, no matter what we did, then perhaps it really does not matter whether we describe the trajectory of an object using equations involving gravitational or inertial frames of reference. Thus our understanding of the principle of equivalence is increased, by having increased the empirical (semantic) content of the principle.

If we were simply given the principle of equivalence, we might not realize that it lets us bring the phenomenon of gravity under the theory of motion that Einstein had pioneered with Special Relativity, which is the reason Einstein needed it. The principle is absolutely crucial, and the thought experiment gives it cognitive significance. Slowly or suddenly, we gain new understanding of a theoretical structure that we may not have otherwise had. And like Maxwell’s, Einstein’s thought experiment creates ties between a mathematical-theoretical principle and a kinaesthetic or operational sense of what we would experience in an imagined scenario if it were true. This increases the empirical content of the principle by allowing us to associate empirical notions like how we would feel, or what we would experience if, say, a crystal vase was liberated in each of Einstein’s elevators.

4 Heisenberg’s Microscope

Heisenberg develops his matrix-mechanical formalization of quantum mechanics in 1925, in “Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen.” 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, 221). The idea that would be explored in Heisenberg’s thought experiment was already present in this revolutionary paper of his, which begins: “It is well known that the formal rules
which are used in quantum theory for calculating observable quantities such as the energy of the hydrogen atom may be seriously criticized on the grounds that they contain, as basic element, relationships between quantities that are apparently unobservable in principle, for example, position and period of revolutions of the electron” (Heisenberg 1925). Notice that Heisenberg’s first words are “It is well known.” Heisenberg was probably not, therefore, trying to produce a thought experiment to establish the uncertainty principle or “the formal rules” of quantum mechanics.

The thought experiment in question 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,” which is partially translated as “On the ‘anschaulichen’ Content of Quantum Theoretical Kinematics and Mechanics.” Hilgevoord (2006) notes that “anschaulichen” has been translated in several ways, so that the title may read “On the Physical Content…” (Wheeler and Zurek 1983), “On the Perceptible Content…” (Blum, Dürr and Rechenberg 1984, listed in the references under Heisenberg 1984), and “On the Perceptual Content…” (Cassidy 1992). Hilgevoord himself considers “intelligible” or “intuitive” as substitutes. Searching for the physical, perceptible, perceptual, intelligible or intuitive content of quantum mechanics seems more in line with a characterization of Heisenberg’s microscope as part of a project of understanding, which I noted at the beginning of this chapter is an umbrella term, rather than knowledge.

The thought experiment is substantially re-worked for Heisenberg’s lecture-tour of the United States in 1929. In these lectures, Heisenberg presents and explains 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, 226). The microscope thought experiment is published in its completed form in 1930, in the book that resulted from these lectures, *The Physical Principles of Quantum Theory*.

As above, Heisenberg’s gamma-ray microscope could be used to convince rival scientists to adopt the Copenhagen interpretation, and therewith, quantum mechanics, by arguing for this interpretation. This could be done either in a weak evidential sense by making it plausible, or by
eliminating defeaters. However, Heisenberg might have also been using the microscope to explore the uncertainty principle itself. Unlike Einstein who used mathematics combined with his elevator thought experiment to substantiate a theoretical intuition, Heisenberg used his thought experiment to explore an experimentally derived regularity. That regularity is expressed in the theoretical vocabulary of quantum mechanics and in the end, is supposed to tell us what the world is like. But it is hard to see right away how it does.

Here is the thought experiment. When we see an everyday object with our eyes, it is because we register visible wavelengths of electromagnetic radiation that bounces off of it. The more precisely we want to determine the position of something by bouncing electromagnetic radiation off of 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: because 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 the precision in measuring the momentum of the particle. This is how Heisenberg uses an imaginary example to work through a theoretical structure which predicts a practical inability to measure two linked quantum observables at the same time, in this case, position and momentum.

To understand the motivation behind the microscope, we require more historical context. Immediately after Heisenberg proved that the matrix and wave mechanical formulations of quantum mechanics were equivalent, Bohr wanted to unite quantum mechanics by producing an interpretation that was consistent with both wave and particle mechanics. This is because Schrödinger was set on the wave mechanical interpretation and Heisenberg on the particle
interpretation. Rather than supporting Bohr, Heisenberg sought an interpretation of quantum mechanics focused on the particle nature of atomic elements that could be visualized. The greatest barrier was the uncertainty principle. Why is it that two elements of a system (like position and momentum) for a particle cannot even in principle be simultaneously measured with arbitrary precision? “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, 105, my emphasis). Heisenberg’s microscope therefore seeks a link between the new theoretical structure and the content, for Heisenberg.

Even 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, 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, 81-85). In other words, the thought experiment gave Heisenberg a way to interpret the new theoretical formalism and the concepts that came with it 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. (2007, 179)

Camilleri’s use of “define” here makes Heisenberg appear to be a conventionalist. If this is true, he is a conventionalist when it comes to establishing connections between theory and experience. Whatever Heisenberg’s epistemology, the microscope seems to be an excellent example of a thought experiment intended to help us come to terms with a new and difficult theoretical structure, whatever other epistemological functions it may have served besides.

The way the thought experiment produces understanding is by connecting our everyday sort of knowledge with the theoretical structures of quantum mechanics. By imagining a microscope that interacts with a high-energy particle (that we imagine as a round solid object) that has recently collided with another moving particle, the thought experiment helps us to see why we should not expect simultaneous measurement of position and momentum with arbitrary precision. Of course, it does this in a misleading way. The reason we cannot simultaneously measure the position and momentum of a quantum particle is not because our apparatuses are not accurate enough to trace post-collision trajectories, it is because of the specifically quantum nature of those systems. I will discuss this failure in more detail below.

5 Darwin’s Vertebrate Eye

The Origin of Species was published in 1859, and we know Darwin was thinking about evolution by natural selection for a long time leading up to that. The eye thought experiment is expanded in each edition, perhaps in response to criticism, but it was with Darwin from the beginning.

Darwin wanted people to understand his system as clearly as possible, since he rightly foresaw empirical, theoretical and theological backlash. Because of Darwin’s careful treatment of
possible objections, especially in Chapter 6 (“Difficulties of the Theory”), his most serious critics did indeed understand the ideas of random mutation, the inheritance of properties, and selection via death or the inability to reproduce. This meant that after the book was published the scientific objections that arose were not due to misunderstanding, but instead pinpointed serious theoretical difficulties (for example, Kelvin on the age of the earth, see Bowler 1983, 23-24; or the “swamping” argument by Fleeming Jenkin, see Jenkin 1897, Bulmer 2004).

In Chapter 6 of the *Origin*, Darwin introduces four main difficulties for his theory. The second asks, “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 1909, 178). When Darwin arrives at this problem, he re-states it as the problem of “organs of extreme perfection and complication” (190), 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. (190)

What exactly does Darwin pretend to find absurd? It is not his theory in general, but the specific application of that theory to very complex cases like the vertebrate eye. How could something so complex 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 (an organism with half-eyes). And at that stage, the half-eye would certainly not have been adaptive for the organism. Eyes only work when they are complete: half an eye has neither purpose nor function. Furthermore half an eye never would have evolved unless it did so specifically in order to produce a complete one later on.

In response, Darwin provides a thought experiment to help us understand his theory, and refute the objection. The thought experiment shows that even in this problematic case, his theory is explanatorily adequate.
We imagine a creature with no eyes. After a mutation, its offspring have a light sensitive cell or patch, somewhere on its body. Surely this would have been adaptive. We imagine this patch growing and shrinking, by additional chance mutations, over a long period of time. A larger light-sensitive patch could register finer differentiations in light levels and thus be more useful than a smaller patch, although only to a degree. Such a patch could enable a creature to tell 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. We can also imagine topological changes to such a light sensitive patch. If it were to become convex, it would collect more light, but it would become vulnerable to damage by sticking out beyond the skin. We can also imagine that patch becoming concave. This would protect it from injury, without limiting its light gathering ability 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 circle, at which time a lens would be very advantageous. Muscles to better focus the light and an increased number of nerves to process it could come later. At each stage of this process, we have an organism with an adaptive mutation. At no point is there a useless half-eye.

The text of the actual thought experiment is less detailed, but the suggested thought experiment is clear. Darwin considers simple eye-like organs possessed by organisms like crustaceans, and gradual changes that could occur which might lead to more complicated organs (Darwin 1909, 191-192). He begins from cells that merely differentiate light from dark, then adds a mechanism to concentrate light rays, and finally an optic nerve capable of reading the light patterns. He adds credibility to the different stages of the thought experiment as he goes through them. He points to the development of the human eye in the womb, citing evidence that it grows not completely but in stages, starting with very simple layers of translucent skin. He also compares the eye to the telescope, which could be built in simple stages from a single lens, just like the eye.

What is the stated 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” (1909, 194). Darwin must therefore convince his reader that he or she will find no such case. He obviously cannot go through all possible cases because there will be many that he will not be
aware of, and there is not enough fossil evidence to satisfy a determined skeptic anyway. Rather he takes a few exemplary cases and shows us how to tackle them by imagining steps that would connect a series of possible stages of development.

This use of the thought experiment might be justificatory in the sense that it refutes a counterargument which denies the possibility of applying Darwin’s theory to all relevant biological systems. There are other ways of understanding Darwin’s thought experiment as justificatory, although they are less convincing.

For instance, we might think that Darwin was actually proposing a specific evolutionary timeline for the vertebrate eye. And indeed, Darwin’s employment of imaginary histories has been criticized by some opponents as “just so stories.” Just so stories in this context are attempts to explain traits by evolutionary accounts that are imaginable, but have no other support (see Lennox 1991, 238, who cites Hull 1973, 319, 339, 342-3; and Kitcher 1985). For example, Rudyard Kipling tells the story of how the leopard got its spots: they were finger-painted on by a friend who understood the benefits of camouflage. This is a nice story, but not a scientific one. As we have seen, Darwin does tell stories for which he does not have evidence. But his stories are not ad hoc because he does not put them forward as being true (incidentally, neither does Kipling). To think that he was, in the complete absence of any fossil evidence, would make Darwin out to be irrationally optimistic about his mental powers. Rather, Darwin was trying to make it clear how his theory and its concepts could be given physical meaning in the case of the eye. This is close to how James Lennox has been characterizing some of Darwin’s well-known thought experiments for some time (1991, 2005).

Lennox characterizes Darwin’s thought experiments as “tests, not of the truth of the statements comprising the theory, but of the explanatory potential of the theory. They are designed either to display, or to challenge, a theory’s ability to explain the full range of phenomena it claims for its domain” (1991, 223). Lennox is advocating something like the understanding-role of thought experiments in science that I have been focused on. He also maintains that these “thought experiments can play [a role] in the evaluation of a theory” (223). This is because theoretical virtues are important, especially during times of scientific revolution, and if Darwin can show
that his theory is indeed capable of explaining what he says it is capable of explaining (displaying the theoretical virtue of explanatory power), this supports the theory.

Lennox sees “Darwinian thought experiments” as justifying Darwin’s theory. I concur that they perform this function. But I also claim that there is another function that these thought experiments perform which concerns understanding. I do not think that Lennox would deny this additional function, since it draws on many of the same facts that his interpretation does. The difference is a difference of emphasis: Lennox is concerned with the “argument patterns [of the Origin] and their relation to the philosophy of science both of Darwin’s time and place and our own,” and more generally, “how it carries conviction”—what makes it a convincing work (1991, 223).

My interpretation of the role of this thought experiment would focus on increasing empirical content instead of displaying explanatory power. Lennox imagines Darwin’s opponents asking whether evolutionary theory can explain all relevant cases. I prefer to characterize Darwin as creating the thought experiment to explain what evolution means, what kind of traits we can expect to see in the world, and how to look for them. I think Darwin is trying to show, just like Maxwell did with his demon, that the dual connection between theory and experience holds. That is, his theoretical structure (evolution by natural selection) is both suggested by experience, and explanatory for experience. On the one hand, Darwin can appeal to embryo stages and the less-developed eyes of other species to use our experience of the world to suggest intermediary stages of ocular evolution. But on the other hand, there is no direct experience we can have of vertebrate ocular evolution. Since we can only see the evolution of the vertebrate eye in our mind, Darwin invents a thought experiment that truncates the passage of time and ignores the millions of failed mutations. Given the thought experiment, we can predict similar evolutionary timelines for each specialized organ, and begin to see all organs as intermediate stages of a dynamic process.

Thought experiments like these make Darwin’s theory understandable, and they are one of the features that make the Origin enjoyable to read even to this day. Darwin shows us how to use the conceptual machinery of his theory by giving examples: he shows us what evolution means. 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 at every stage, we will by the end of Chapter 6 of the *Origin*. So I agree with Lennox that the thought experiment displays the explanatory power of evolution by natural selection, but I argue that it does after it has also rendered the theory more understandable.

6 Understanding Theoretical Structures

I have provided historical evidence that the role of some important thought experiments is not merely one of providing evidence or generating knowledge, as one might have supposed given their role in scientific revolutions. Instead, there is a richer and more complicated relationship between thought experiments and theory, and in every case examined, the production of understanding is a principal function. Whatever other epistemological purposes they may serve, these thought experiments provide a way of coping with theoretical structures. By “theoretical structures,” I mean the models, laws, concepts and relations between them that are among the elements constitutive of scientific theories. In scientific revolutions, these are usually among the first items proposed, and their growing importance motivates scientists and students to understand them.

As we have seen, the thought experiments discussed above are in a sense independent of the theoretical structures they pertain to, since they can appear before, simultaneous with, or after those structures are fully developed and justified. For instance, Einstein’s elevator appears before the equivalence principle. Another example I did not discuss but which perhaps fits this description is Einstein’s chasing the lightwave. Einstein reflects, “If one runs after a light wave with light velocity, then one would encounter a time-independent wavefield. However, something like that does not seem to exist! This was the first juvenile thought experiment which has to do with the special theory of relativity.” Later, he claims “after ten years of reflection such a principle [Special Relativity] resulted from [this] paradox upon which I had already hit at the age of sixteen” (Einstein 1951, 52; and see Pais 1982, 132).

A thought experiment that appears concurrently with the new structure is Darwin’s eye. This relationship might obtain when the author of a new structure foresees difficulty understanding it, and either anticipates an answer, or provides one that they themselves find useful.
Finally, a thought experiment may be developed after the relevant structure is presented to the scientific community, as with Maxwell’s demon, the clock in the box, and Heisenberg’s microscope. I would think that when a new structure makes possible accurate empirical predictions and relatively easy to use, but carries with it new concepts, types of conceptual relations, or ways of thinking, we should expect a thought experiment (or something like it) to follow its introduction. A simple example is Einstein’s train, which was used to introduce the theory of Special Relativity. As Pais recalls, “in the early days it was easier to understand the mathematics of Special Relativity than the physics…it was not a simple matter to assimilate a new kinematics as a lasting substitute for the old aether dynamics” (1982, 115). What Einstein (and Lorentz and Poincaré) gave us was a better way to deal with electromagnetic phenomena. Very soon after, however, we needed thought experiments like Einstein’s train to tell us what it meant for concepts like SIMULTANEITY.

All of this suggests that the interpretation of theoretical structures is in fact a part of the process of science. But is the need to interpret things a mere psychological quirk that we could in principle dispense with, or is interpretation necessary for achieving the aims of science? To argue for the latter option, it seems that in some cases we need to interpret or understand a theoretical structure in order to make any empirical use of it at all. For instance, without the kind of understanding that Darwin’s eye provides, we might not know how to apply evolution by natural selection to complex organs. And once we have been exposed to the thought experiment, we can adopt its methodology and make predictions about the intervening stages of evolution for other organs (or traits, or species). In this way Darwin presents not merely a piece of support for his theory or a way to increase understanding, but a way of doing science. In other cases, the empirical import of a new theoretical structure may be clearer, and no thought experiment will be needed; interpretation takes place another way. Yet even here a thought experiment may be useful if it reveals additional empirical consequences or if it makes easier the cognitive assimilation of the new empirical import.

Finally, what do I mean when I say that the thought experiments discussed above are ways of “coping with” or “dealing with” theoretical structures? I mean that they make such structures cognitively manageable for humans. This is necessary because new theoretical structures on their own do not make it clear what their substance or semantic content is, no matter how many
derivations we perform, observations we make, experiments we conduct, or computer models we run. This is not to maintain a strict distinction between form and content: I allow that even our most abstract theoretical structures always have some empirical content. This does not change the fact that when we are presented with a theoretical structure that breaks sharply from what existed previously, as with some of the concepts or relations in relativity theory and quantum mechanics, we find ourselves in the uncomfortable position of having an empirically adequate theory whose empirical relevance or substance we do not fully understand. For example, Newton could describe the behaviour of objects using his universal theory of gravitation, but he could not say what gravity was. Newton was uncomfortable with this, and his opponents capitalized on it. Similar scenarios can be found throughout the history of science, for example, Boyle and the “spring” of the air (Shapin 1996, 104), and Einstein and the light quantum (Howard forthcoming, 6). If a candidate for scientific knowledge is something like a list of propositions that we believe are justified and true, and scientific understanding is a grasp of the kind of world such a list describes and how to navigate it, then knowledge and understanding (of the same thing) can increase at different rates. We may have understanding of something that we have not yet justified, or we might have knowledge that we do not yet understand.

I do not want to imply that thought experiments themselves are necessary for scientific progress, only that the understanding they are capable of creating is. This idea needs substantiation. To anticipate later chapters I will discuss understanding in a few words now. If scientific progress is the increase of scientific knowledge, then the question of the relation between understanding and progress becomes one of the relation between understanding and knowledge. According to Peter Kosso: “knowledge of many facts does not amount to understanding unless one also has a sense of how the facts fit together” (2006, 173). He invites us to remember the Omniscienter from Pierre Dumal’s novel A Night of Serious Drinking, over whose chair it reads, “I know everything, but I do not understand any of it.” Kosso suspects that the Omniscienter has spent too much time gathering evidence and too little time thinking about it. He has taken the piecemeal empiricism too seriously and overwhelmed his science with observation. Too many data have left too little room for understanding. There are examples of knowledge without understanding in the physical sciences, and
they are found in the most empirically dependent sciences or in any science at the time of new empirical discovery. (182)

Without understanding, we are left with tables and lists of empirical data that are disconnected from human aims and practices. More data will not remedy this. Scientific understanding then, is related to the meaning and relevance of scientific data for scientists and society, and the ability to apply that data broadly. This is why Steven Weinberg remarks that General Relativity offers more understanding than quantum mechanics, because the latter cannot as easily be bridged to our other stores of knowledge (Weinberg 1992, Kosso 2006, 184). If it is true both that we need to understand our theories, and that quantum mechanics is especially difficult to understand, then we should expect a great deal of thought experiments in quantum mechanics, especially when the theory was introduced. And indeed, it is probably the period most replete with thought experiments in the history of science.

The rest of this dissertation will be split between providing evidence for the idea that thought experiments can produce or increase scientific understanding in this sense, and explaining how thought experiments might do this. The exact role thought experiments play and how they play it is not equal in all the above cases, but they are all plausibly construed as increasing the subjective semantic content of a theoretical structure. Einstein’s train, for example, explains new uses for concepts like SIMULTANEITY. Einstein’s elevator and Maxwell’s demon connect theoretical structures to possible but non-actual experiences. Others connect new structures to old ones, as Bohr’s version of the clock in the box connects Heisenberg’s time-energy uncertainty relation to relativity theory. And I will argue that what lies at the bottom of each increase in understanding is the bridging between theory on the one hand, and experience, knowledge or abilities on the other. The nature of the bridging function will be addressed in detail in Chapter 7.

In the next chapter, I will substantiate the claim that increasing understanding is a goal of scientists by providing evidence from cognitive and social science studies on working scientists and students of science. I will show that thought experiments are daily tools of the science worker, educator and student, and that they employ these tools predominantly to increase understanding.
Chapter 6
Scientific Imagination and the Problem of Coordination

Knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution. It is, strictly speaking, a real factor in scientific research.

(Einstein 1931, 97)

1 The Imagination

Blaise Pascal called the imagination that “deceitful part in man, that mistress of error and falsity.” He said it was “all-powerful,” and the “enemy of reason” (1901, 51). Malebranche referred to the imagination as “a fool who likes to play the fool” (1842, 79), and many fictional and historical catastrophes can indeed cite over-active imaginations at their root. It is imagination that leads Goethe’s Werther to his infamous sorrows, and imagination is behind the tragic ambition of Mr. and Ms. Macbeth. Mein Kampf provides a more terrifying instance. According to George Orwell, Hitler saw himself as “the martyr, the victim, Prometheus chained to the rock, the self-sacrificing hero who fights single-handed against impossible odds. If he were killing a mouse he would know how to make it seem like a dragon” (1940, 14).

Yet the imagination is also a positive force in the human condition. To reverse the sexism of Pascal, we are lucky to have imagination. Without her we might have no values, no ethics, no understanding, and no knowledge. Tribute is regularly paid to the imagination in personal histories of science. Francis Jacob, a Nobel Prize winning biologist, has written:

It was not a simple accumulation of facts that led Newton, in his mother’s garden one day, suddenly to see the moon as a ball thrown far enough to fall exactly at the speed of the horizon, all around the earth. Or that led Planck to compare the radiation of heat to a hail of quanta. Or William Harvey to see in the bared heart of a fish the thudding of a mechanical pump. In each case they perceived an analogy unnoticed up till then. As Arthur Koestler pointed out, everything in this way of thinking seems different from that
of King Solomon when he compares the beasts of his beloved Shulamite to a pair of fawns, or that of Shakespeare’s Macbeth, when he sees life as ‘a tale told by an idiot, full of sound and fury.’ And yet, despite the very different means of expression used by the poet and the scientist, imagination works in the same way. It is often the idea of a new metaphor that guides the scientist. An object, an event, is suddenly perceived in an unusual and revealing light, as if someone abruptly tore off a veil that, till then, had covered our eyes. (2001, 119)

Jacob reminds us that no agglomeration of facts can give us the power over nature that science seeks, or the beauty and novelty of art. Dustin Stokes (2014) argues similarly that even if Bach had known all there was to know about musical relationships, this would not have been sufficient to compose The Well-Tempered Clavier (159-160). This echoes Jacob’s claim above: whatever is going on in scientific discovery, it is not merely the collection of facts. Other Nobel Prize winning scientists gesture to similar senses of imaginative artistry and its necessity in science (for example Einstein 1931, 97; 1934, 163; and see Holton 1996, Hadamard 1996).

It was common in the philosophy of science for a long time to hold that the imagination was not relevant for scientific progress other than in the context of discovery. Its most ardent opponents were positivism (see De Regt 2009) and behaviourism (see Barsalou 1999). The main complaint was that the imagination is subjective; what you can imagine is not necessarily what I can imagine. And what is subjective has no place with the objective elements of science like experiment, theory and knowledge. Partially thanks to the influence of science studies since the 1960s, many philosophers and cognitive scientists have given up the distinction between the contexts of discovery and justification and therefore should allow a role for subjective elements in objective science. Another reason for renewed interest in the imagination is the role of the imagination in scientific thought experiments. In his introduction to a special issue of Metaphilosophy on thought experiments, René van Woudenberg claims that “the imagination, perhaps surprisingly, plays an important role in the process of obtaining knowledge: knowledge of certain normative issues, of possibilities, of moral truths, of certain physical matters, of one’s self, and more” (van Woudenberg 2006, 160).
This chapter will explore the epistemological role of imaginary experiments in science developed in the last chapter by considering recent results from social and cognitive science.

A preliminary obstacle is that cognitive science studies rarely refer to the imagination in a general way. Instead they speak of mental images or the creation of analogies or mental models. Something similar goes on in mainstream analytic philosophy, which deals with the imagination as something that tests modal propositions by seeing whether they are conceivable, or produces psychological states which obey special norms, and much else (see Gendler 2013 for a sample of characterizations of imagination). In order to connect empirical and epistemological issues, then, I maintain a very inclusive reading of imagination, delimiting not much more than the mental ability to (re)present things on purpose to ourselves that are not now present via the senses. These (re)presentations need not be propositional or static (like images), and to allow space for rationalism, their content need not consist entirely of permutations of previous experience. Finally, they need not depart from the truth. Imagining a Boeing 747 at the bottom of the Mariana trench is no less an imagining if there is in fact a Boeing 747 there.

Let us turn then, to results in cognitive science. Şule Kösem and Ömer Özdemir have recently claimed that “imagery is an indispensable component of scientific reasoning” (2014, 887), and many others agree (including Brown 2006; Clement 1993, 2008, 2009; Gilbert and Reiner 2000; Klassen 2006; Lattery 2001; Reiner and Burko 2003; Reiner and Gilbert 2004). Add to this the claim of Gooding (1992, 285) that “visual perception is crucial because the ability to visualize is necessary to most if not all thought experiments.” Still, it is not immediately obvious how we should go about investigating this component of reasoning. In the last chapter I considered cases in the history of science. Lynn Stephens and John Clement argue that while such an exercise may be helpful, it is not enough to discern the cognitive mechanisms that underlie imagistic mental reasoning of the type we find in scientific thought experiments. They write:

It is difficult to analyse the mental processes that allow a scientist to generate and run a thought experiment during an investigation by using historical data because the original thought process can easily be buried under many changes and refinements the author carries out before publishing a thought experiment. Also, for many thought experiments it
is hard to know whether they were originally part of a discovery process or created after the investigation to convince others. (2012, 160)

Historical details like those presented in Chapter 5 can only take us so far. We must also study thinking agents in real-time. What follows is a survey of all the major cognitive science studies concerning thought experiments in science (or science education) in the last several years. First I will look at studies done on thought experiments in science education, and then in scientific problem-solving.

Reiner and Gilbert (2000) discuss thought experiments that appear in textbooks. They begin by cataloguing which thought experiments appear where, and for what purpose. Then they compare the original and textbook presentations of famous thought experiments, concluding that thought experiments help students and scientists understand scientific concepts. What does it mean for a thought experiment to help us understand something? The authors cite Stephen Toulmin (1972) in explicating what they mean by understanding a concept in terms of being able to use it. A concept of any kind is capable of use, and therefore understood, if two criteria are met: if it is intelligible, in that the user knows what it means; and if it is fruitful, in that it enables the user to achieve a goal or to identify new possibilities.

Most of the studies to be discussed are easily brought under this rubric. Scientists and their students must be able to use new concepts, otherwise they serve no purpose. And they cannot use a new concept without knowing what it means, or in other words, without the concept being made intelligible. Intelligibility is not always so easy to achieve, and we will see that thought experiments are sometimes capable of affording it. Secondly, if one understands a concept, one can achieve something with it. Thought experiments help us explore the consequences of adopting certain viewpoints, and to see how conceptualized phenomena interrelate, and this opens up new possibilities for theorizing, modelling, and constructing experiments. We will sharpen these ideas of understanding, intelligibility and usability later on.

Building on this framework, Reiner and Gilbert argue that thought experiments in science textbooks are not used optimally. In scientific literature, most thought experiments are presented in the following way: we begin with a scenario or statement of some problem. We create an imaginary world to help us explore the scenario or problem. We “set up” or “design” a thought
experiment in this imaginary world, which we then “run” and “observe.” Finally, we draw a conclusion about the initial problem or scenario. This presentation-style spurs those who follow its twists and turns to make new connections on their own. On the other hand, textbook presentations are only “mental simulations.” Here the conclusion of the thought experiment is given first, and then the imagined scenario is introduced, which lends credence to the conclusion. In a mental simulation, students do not vary variables in their minds; they simply follow along with a text (Reiner and Gilbert 2000). This practice is less likely to achieve Toulmin’s two conditions, because if you do not perform the thought experiment or otherwise establish the conceptual connections for yourself, a scientific concept or theoretical structure will have diminished meaning for you, or no meaning at all. And in this case, you will not be able to see how best to use that structure.

Velentzas, Halkia and Skordoulis (2007) look at textbooks as well, and they show that what Brown calls “constructive” thought experiments (Brown 1991a, 36), that is, those thought experiments that provide evidence for or establish a theory, are preferred by textbook authors to what Brown calls “destructive” thought experiments (Brown 1991a, 34), which function as counterexamples. The thought experiments used most commonly in physics textbooks are Einstein’s train, Einstein’s Elevator, and Heisenberg’s Microscope. Perhaps this is because constructive thought experiments show students how their everyday experiences can translate into modern day physical theory, which motivates them to learn (2007, 365ff.). In other words, they “build bridges between students’ knowledge and everyday experience and the new or modified concepts and principles which have to be learned” (359). Building such bridges would certainly help to make a new concept intelligible and useable for a student. One way to interpret the results of this study is that destructive thought experiments are not as useful for building such bridges. However we will see below that destructive thought experiments are indeed used successfully for this purpose, and therefore something else must be responsible for the popularity of constructive thought experiments in science textbooks. In any case, their conclusion concerning constructive thought experiments accords nicely with the results of the historical studies in Chapter 5, above.

This study inspires several more. In (Velentzas and Halkia 2011), the authors discuss Heisenberg’s Microscope “as an example of using thought experiments in teaching physics
theories to students.” They begin by paraphrasing Alexander Koyré, who claims that thought experiments “help scientists to bridge the gap between empirical facts and theoretical concepts” (2011, 525; from Koyré 1968). They sympathize, and argue that while Heisenberg’s microscope is not generally well regarded by physicists (either at Heisenberg’s time or now), the thought experiment is still effective for introducing the uncertainty principle in quantum mechanics, which Velentzas and Halkia teach to 40 high school students in grade 11 using the thought experiment. Here is what they did. First, some fundamentals of quantum mechanics were introduced in the classroom. Then students were given the imaginary scenario and allowed to work through their guesses about what would happen while being directed through a series of leading questions. If they went too far off track, they were guided back. Velentzas and Halkia recorded the sessions in order to code and analyze them, and administered a test two weeks later for comprehension. They concluded that many students could and did learn the uncertainty principle using the thought experiment. And not merely for the case of gamma rays and microscopes; they appreciated the principle independently of any considerations of specific measuring apparatuses.

Next they turned to General Relativity (Velentzas and Halkia 2013a). In this paper the authors concluded that thought experiments in General Relativity make it possible for students to “grasp physical laws and principles which demand a high degree of abstract thinking, such as the principle of equivalence and the consequences of the constancy of the speed of light to concepts of time and space” (2013a, 3026). They found this achievement more surprising than in the case of the uncertainty principle, because these students had strong folk intuitions which interfered with understanding General Relativity. For example, students had difficulty with INERTIA, imbuing what looks like magical power to it. Students also assumed that their intuitive concept of time could not be wrong; that space is empty and separate from time; that observers may have a point of view, but that this has no bearing on physical laws since there is always a more objective point of view that we can escape to which makes clear the “real” answer (Arriassecq and Greca 2012).

However, the authors conveyed the concepts of General Relativity to the students successfully, letting them work through thought experiments like Einstein’s elevator and train. They recorded the sessions and analyzed them, and then administered a test two weeks later for comprehension.
They conclude that thought experiments are used “both for clarifying the consequences of physics theories and for bridging the gap between the abstract concepts inherent in the theories and everyday life experiences” (3027).

Finally, in (Velentzas and Halkia 2013b) the authors turned to Newton’s cannon. As in the above two cases, the authors asked the students to work through the thought experiment on their own and see that projectile motion and orbital motion are governed by the same laws. The authors claim that Newton’s thought experiment “can act as a bridge which enables students to correlate the idea of the ‘downward’ motion of objects drawn from their everyday experience with the same objects’ motion ‘to the center of the Earth’” (2623). To make this possible, students had to see the earth from above, and extend their experience with regular projectile motion to a scale large enough to include suborbital and orbital motion. This allowed them “to link the motion of a projectile as it can be observed in everyday situations with the possible case of a projectile that can move continuously parallel to the ground in a context where the whole Earth is visible” (2623).

Notice that the metaphor of “bridging” is common to all of these studies, and will continue to be invoked below. It is significant because it relates to both of Toulmin’s conditions. Bridging theoretical structures to everyday experience and abilities increases the intelligibility of a concept because such connections allow us to import existing semantic content and understanding into a new theoretical structure. When we increase semantic content we know more about what something means, which is another way of saying that it is intelligible to us. And when a bridge opens, new territory becomes accessible. The territory was already there, but we did not have the ability to access it. This new ability, and those that come along with it, increase the fruitfulness of a concept. For instance we saw in the last chapter how Darwin’s eye thought experiment provides us with a way of using the theoretical structures of evolutionary biology, by showing us a new way to conceive of trait evolution.

Velentzas and Halkia (2013b) conclude that thought experiments are useful in science education because they help students learn to apply difficult scientific concepts. But there are two other interesting conclusions they draw. One is that thought experiments are pedagogically superior to computer simulations, because only in a thought experiment is it completely up to the student to
see how the outcome of the imagined scenario results from the parameters. A computer simulation that portrays the earth from above and allows the student to program in different projectile velocities and see how these changes affect the motion of a projectile will be useful and probably more illuminating than calculating consequences of Newton’s laws. But in this case the student takes a passive role by setting the parameters and waiting to see what happens. In a thought experiment, students mentally “set” the parameters, and then in addition have to figure out what will happen. Instead of trusting to the algorithms of a computer, students must justify their mental work and provide some reason to believe the system will evolve as it does in their imagination. Also, talking through imaginary scenarios enables teachers to see where a student stands with respect to their comprehension of the theory. Therefore the authors conclude that thought experiments will not be replaced by computer simulations in the near future, at least, not in the classroom.

Their second conclusion is also noteworthy, namely, that “in any experiment, the manipulation of ideas is more important than the manipulation of materials” (Velentzas and Halkia 2013b, 2638). That is, “hands on is less important as compared to minds on” (Duit and Tesch 2010). Presumably the authors mean that manipulating laboratory equipment will not be informative for a student who does not grasp the meaning of those events. And with respect to the goal of increasing scientific understanding, this is something worth stressing.

Now I want to look a little closer at how thought experiments originate in situ.

In “The Symbiotic Roles of Empirical Experimentation and Thought Experimentation in the Learning of Physics,” Miriam Reiner and John Gilbert argue that in the course of solving empirical problems, subjects often construct and run thought experiments spontaneously. They conclude that “the process of alternating between these two modes—empirically experimenting and experimenting in thought—leads towards a convergence on scientifically acceptable concepts” (2004, 1819). In other words, thought and empirical experiments appear in conjunction, and this is for the best, because together they enable us to go from “seeing” a physical phenomenon to “knowing” about it (1820).

To reach this conclusion, Reiner and Gilbert asked students to analyze a physical mechanism that behaved in an unexpected way. Two heavy wheels were set into a base, side by side, and each
was free to spin. If one is made to spin quickly, the other does nothing. But as it slows down, the other begins to spin and speed up, until the first comes to a complete stop. When the second wheel slows down, the first starts spinning again. The reason for this behaviour is a set of hidden magnets contained in the wheels. Given a list of the materials out of which the mechanism was built, the subjects were asked to figure out what was going on. Each participant naturally adopted approximately the same methodology. He or she began by identifying the various physical mechanisms in a general way using concepts like force, acceleration, weight, direction, and so on. These variables were combined to construct (mental or physical) models that attempt to capture what is observed in the mechanism. These models were abstracted further into what the authors called a “representational space,” where the relationships between features of the mechanism are represented, often with the help of pen and paper. Finally, the students created and used imaginary worlds to test their models through thought experiments.

The authors claim that this is an instance of the move from perception to knowledge, and “how concepts emerge out of touching and seeing. [This move] forges links between the bodily and the mental, between the physical and the cognitive, faculties.” The authors break up the process into four main steps. First, “sensory interaction with the physical set-up in the experiment triggers knowledge associated with memories of sensations.” Second, “knowledge associated with sensations is spontaneously represented in pictorial representations.” Third, those “representations are used for communication, their meaning shared across situations, across subjects, used for conceptual negotiation.” And finally, the “representations are used for running experiments in thought. The results of these thought experiments are used to refine the set-up of the physical experiment” (2004, 1831).

Note the cyclic nature of the process. Perception, memory and tacit knowledge feed into the creation of new representations, which are communicated and negotiated using models and thought experiments and then tested. But the results of the tests cause more representations to be created, which feed into more models and more tests. Experience does not justify the results of the imaginary experiments any more than the imagination justifies the results of physical modelling or laboratory experiments. The two cognitive achievements are symbiotic.
There are three more papers I want to examine before I draw conclusions. The first is Kösem and Özdemir (2014). This study asks what thought experiments are used for. To answer they collected three groups of subjects, each with a different level of expertise in physics. They produced trials to find out what types of thought experiments would spontaneously be invented, and for what purposes.

The problems they asked were drawn from dynamics or mechanics. The first group consisted of doctoral graduates, the second university undergraduates, and the third high school students in grade 12. The total number of thought experiments invented by each of the three groups was roughly equal.

In terms of the type of the thought experiments invented, each student either modified an object in an imaginary scenario (for example, the size of a car), or a variable (for example, its velocity). When they modified the object, they did so either to match a familiar case, like the way they felt when riding a bus, or a simpler case, for example, by dissecting a problem into several smaller, easier problems. When they modified a variable, they either eliminated or minimized the value of the variable, which helped them focus on the relationships among other variables, or they increased the value of a variable to make its effect on the system more obvious.

In general, changing the problem to a more familiar case by modifying the object was the most common strategy for the undergraduate and high school groups. Modifying the variables was done most often by the doctoral group, and seldom by the others.

In terms of the aim of the thought experiments, there were several. Sometimes a subject would have an intuition, which they explored using a thought experiment. At other times, a subject might have an independent reason for believing something, which they chose to illuminate with a thought experiment while trying to report or justify it. Still other times the thought experiment played the role of a proof. The undergraduates used thought experiments as a proof more than the other groups. The high school students and doctoral graduates very rarely used thought experiments this way. Across all three groups, however, by far, “the most frequently observed purpose of using a thought experiment is for ‘explanation’” (882). That is, “to communicate ideas, or exemplify the solution” (879).
Finally, there are two well-known studies of expert scientists working in vivo that I should include. First, Trafton, Trickett and Mintz (2005) show that scientists use the imagination to manipulate mental representations. They argue that scientists create what Clement later calls “overlay simulations” (2009) between external and mental representations. That is, scientists compare and align mental and external representations, checking for fit or feature-similarity. The authors found that the scientists performed this sort of mental manipulation of spatial representations more often in their heads than they did using computer screens, even when visualization software is available (2005, 97).

In a second study, Trickett and Trafton (2007) build on these results, arguing that scientists spontaneously invent “small-scale” or “local” thought experiments (867) in times of “informational uncertainty” (843). Scientists perform thought experiments in such conditions to “develop a general, or high-level, understanding of a system” (844). The authors focus on the data analysis phase of research, in which scientists must negotiate uncertainty to see what information the data presents, and interpret it. Employing “what if” reasoning helps scientists test out alternate interpretations of their facts, fill in holes in their data, and see how the data fits with existing research questions and background theories. Trickett and Trafton predict that thought experiments “will be used by experts when they are working either outside their immediate area of expertise or on their own cutting edge research—that is, in situations that go beyond the limits of their current knowledge” (867).

To summarize: in almost every one of these articles, one of the main conclusions is that thought experiments are important because they bridge conceptual/theoretical knowledge to kinaesthetic or kinematic knowledge. And even when this is not an explicit conclusion, all of the above results are consistent with it. What does this tell us about the epistemological role of thought experiments in science? If we separate the action of bridging existing instances of knowledge from the action of creating new instances of knowledge, we see that thought experiments are often instances of the first kind of action, whether or not they can also be properly characterized as instances of the second. And the thought experiments we examined in this chapter were more often used to explore or interpret conceptual solutions to problems, communicate ideas, or model scenarios, than they were to provide concrete solutions to problems. That is to say, they increase understanding by exploration rather than (or in addition to) producing new knowledge. In fact,
Özdemir (2009) argues that students learn to shy away from using thought experiments as evidence in physics as they mature, although they do not shy away from using them to communicate and explore. It is certainly possible that this trend could be maintained in the professional careers of scientists everywhere. If so, this provides even more reason to focus on the exploratory and semantic uses of thought experiments.

Furthermore, all of the above studies produce results that support the idea that thought experiments create understanding in at least one of Toulmin’s two ways. Velentzas and Halkia showed in the course of three studies that students used thought experiments to bridge empirical knowledge and theoretical structures. Gilbert and Reiner saw a symbiotic relationship between thought and laboratory experiments, which were performed in a way that “negotiated concepts” through communication and exploration, which made a student’s concepts and models intelligible to him or herself, and also to his or her fellows. Stephens and Clement argue that thought experiments “appear to have considerable value as a sense-making strategy” (2006, 1). Kösem and Özdemir found that the most common use of thought experiments across different groups was to “communicate ideas or exemplify a solution.” Trafton, Trickett and Mintz found scientists employing thought experiments to compare, align and manipulate representations, especially for communication. In all these cases, it has been emphasized that if we do not make sense of a theoretical structure, we cannot use it. This is precisely the conclusion I reached at the end of the last chapter after considering historical case studies.

Let us turn to some philosophical considerations of these results. First, are these less-than-knowledge producing roles that we have just discussed epistemological? I think the answer is clearly: Yes. If these roles produce understanding in addition to knowledge, we are able to draw on the quickly expanding resources in the philosophy of understanding, a subfield of epistemology. Understanding was of course rejected as a topic of serious study in the philosophy of science around the time of the logical positivists, because it was associated with a psychological and subjective feeling (especially by Hempel, see De Regt et al. 2009, 3-5; De Regt 2009, 22-24). This feeling might be part of good science and provide clues concerning what to research next (Lipton 2009, Grimm 2009, Thagard and Stewart 2011), but it might also be categorically irrelevant and even misleading (see Ylikoski 2009).
These days, while we lack a unanimously accepted definition in the literature, philosophers do mean something more robust than a mere feeling. In this sense, “understanding” is like “thought experiment”—a vague but useful term. (Examples of each can be produced at will). Still, philosophers might agree on a few things. Understanding emerges from the combination of a subject, an object, and a third thing, which can be a model, an experiment, a theory, a thought experiment or an explanation. One way to find out if we understand something is to see what we can do: when we understand something, we can use it, either in conversation or to achieve some end. Another way to test for understanding is to see if we can explain it “in our own words,” that is, using concepts, structures and knowledge we already have. We have all administered and passed such tests countless times, which I take as evidence that understanding is desirable. By understanding something fundamental and important, we can come to see the world in a new way. Advances like these may be associated with an “aha!” moment, but they are much more than that: they are genuine cognitive achievements. This is why I continuously return to Toulmin’s characterization of what it takes to be able to use a concept. In Chapter 5 we saw that in the history of science, many thought experiments have made theoretical structure(s) intelligible and usable, and have become famous for providing understanding in the above senses.

There are three questions that arise given these claims about thought experiments and their role in science. First, given that scientists do pursue understanding, we should like to know if they must. That is, what happens in science when we do not understand our theoretical structures? Second, how do thought experiments provide understanding? And finally, what is understanding, exactly?

The first question will be addressed immediately after this paragraph. Some hints of an answer to the second question have been given above, but a general account is desirable. Chapter 7 sketches the outline of such an account. The final and most general question will be considered in the last half of Chapter 7.

I argue for the strong position that scientists must pursue understanding to make scientific progress. For one, this is because increasing understanding via thought experiments solves what has been called the problem of coordination, which must be solved over and over again for
science to proceed. The problem of coordination is what Maxwell, Einstein, Bohr, Heisenberg, Darwin, and the study participants of this chapter were trying to overcome with their thought experiments. In the next section I will outline the problem, and quickly summarize some philosophical solutions. I contend that the specific version of the problem of coordination that applies to theoretical structures in general has not been solved, although there are important lessons to be learned from the literature. A general solution is sketched in Chapter 7, which explains how thought experiments might overcome this problem and thereby increase understanding.

2 The Problem of Coordination

I have suggested that thought experiments increase understanding by causing in us a kind of bridging action that opens new epistemological avenues by connecting theoretical structures to experience and abilities. This section is about that connecting action and about what gets connected by thought experiments. I will show that this action is a noteworthy epistemic achievement by arguing that the problem of establishing or increasing the empirical content of theoretical structures via forging new connections is an instance of a solution to what has recently been discussed as the problem of coordination (van Fraassen 2008, 115). First, therefore, I will establish that the problem of coordination is a real obstacle to the progress of science, and second, that the increased understanding provided by thought experiments solves it.

The problem of coordination concerns the “laws, principles, or equations that involve terms specific to [a scientific] theory” (van Fraassen 2008, 115). These laws, principles and equations are what I have been calling theoretical structures. van Fraassen’s examples include $F = ma$ and $PV = nRt$. These structures cite terms such as distance, velocity, time, pressure and temperature, which provide some empirical content to the relevant theoretical structures. But that empirical content is incomplete without some understanding of the link that connects such terms. Is force the exact same thing as mass multiplied by acceleration, and if not, what does that equals sign mean? If it is purely numerical, how do we measure the quantities? What causes them to be related in this way?

The problem of coordination is usually discussed with respect to measurement, because what TEMPERATURE refers to is thought to be a question answerable in metrology. But while
thermometers hopefully provide part of the content for TEMPORATURE, it is hard to say what plays that role for the content for $E=mc^2$. This is where I think thought experiments can fit in. As proposed solutions to the problem of coordination concern measurement, they will not solve the generalized problem that I want to address, as many theoretical structures cannot be given empirical content via the assigning of numbers to variables. However some of the solutions offered will prove helpful.

The term “coordination” in the relevant sense was coined by Ernst Mach to describe what must happen between a theory’s variables and the things in the world referred to by the theory. To measure the temperature of a room, for example, we need a thermometer, which measures a state of heat. To operate a thermometer you need some kind of index assigning numbers to states. In modern thermometers this assignment will be related to the volume of a mass of mercury at constant pressure, which depends on indices for length and volume change. If we suppose that we have such indices, and a measuring device which is calibrated to them, we must still make reference to the thing to measure. This is the state of heat of the room. But what is that? What is the real thing (not the words or concepts) that corresponds to THE STATE OF HEAT OF THE ROOM? The problem does not go away if we substitute AVERAGE KINETIC ENERGY OF THE PARTICLES IN A SPECIFIED VOLUME for HEAT, because in that case we still need to find the real thing that this more complicated concept refers to, and so on for each of the other terms in that phrase.

Reichenbach summarizes the problem in the following way. Unlike mathematical objects,

the physical object cannot be determined by axioms and definitions. It is a thing of the real world, not an object of the logical world of mathematics. Offhand it looks as if the method of representing physical events by mathematical equations is the same as that of mathematics. Physics has developed the method of defining one magnitude in terms of others by relating them to more and more general magnitudes and by ultimately arriving at “axioms,” that is, the fundamental equations of physics. Yet what is obtained in this fashion is just a system of mathematical relations. What is lacking in such systems is a statement regarding the significance of physics, the assertion that the system of equations is true for reality. (1920, 36)
Reichenbach’s point is that drawing connections between concepts is not good enough to connect them to the world, because those connections themselves are merely more concepts.

In a talk called “Geometry and Experience,” Einstein recognizes this situation as well, and like Mach, he uses the word “coordination” to describe it. He writes, “It is clear that the system of concepts of axiomatic geometry alone cannot make any assertion as to the behaviour of those objects of reality which we designate as practically rigid bodies. To be able to make such assertions, geometry must be stripped of its merely logical-formal character by the coordination of experiencable objects of reality with the empty conceptual schemata of axiomatic geometry” (Einstein 1921). The problem is one about the connection between mathematical-theoretical structures and the world to which they are supposed to apply. Einstein also spoke of the “ever-widening logical gap between the basic concepts and laws on one side and the consequences to be correlated with our experiences on the other—a gap which widens progressively with the developing unification of the logical structure, that is with the reduction in the number of the logically independent conceptual elements required for the basis of the whole system” (1934, 165). In other words, he recognized the need for something to bridge the gap between our theoretical structures, including laws, concepts, equations and models, and what they say about the world. Further, he considers the possibility that as physics becomes more refined and unified, it must make use of more and more abstract notions and relations to accommodate all its information by interrelating it. When this happens, the gap between theory and experience grows wider.

van Fraassen summarizes: “coordination…is required for the new theoretical assertions to have any empirical content at all” (2007, 117). This goes too far if van Fraassen is claiming that without an intentional act of coordination, every theoretical structure of science will be completely void of content. Most new theoretical structures bring with them the empirical content of their constituent or related concepts. And often these structures are models which function by analogy or direct representation which make their content easier to grasp. In other words, the problem of coordination will not always be difficult to solve, since coordinated structures import their content into other structures which contain or relate them. This, however, does not mean that the resulting structures, even those containing components all of which have been previously coordinated, will not need to be coordinated. This is because the new
relationships between the parts of the structure can be difficult to understand. **Space** and **time** are arguably among the first concepts coordinated by humans, whether they realize it or not. General Relativity combines them into **spacetime**. Despite the appearance of a simple conceptual combination (like **red shirt**), it not immediately obvious how to connect **spacetime** with experience or physical phenomena, and so we need some imaginative tool, like a thought experiment. Or, at least, that is what I will argue.

One might object that Mach, Reichenbach, Einstein and others only saw a problem here because of the philosophical environment of their time. Developments in philosophy of science about the ontological nature of theories have already eliminated the problem. The old view of scientific theory was the “syntactic view,” according to which theories are nothing more than a conjunction of the axiomatized sentences the theory takes to be true. According to this view, the problem of coordination is especially pertinent because we need a way to relate the theoretical conjunction to experience, which is not theoretical. One solution is “bridging principles,” as in Carnap (1950). This conception was famously rejected (by Quine) for the “semantic view,” according to which a theory is just the set of models used by the theory. Here, theory connects to experience through its models, which are mediating entities.

However the problem of coordination persists even after adopting the semantic view, because it is always sensible to ask what a given model is a model of. This is true no matter how we conceive of the representation relations that obtain between model and target. Models refer, surely, but the model itself (considered as an abstract non-linguistic structure) does not tell us what it refers to, or how something abstract matches up with the world, which is not abstract. In this respect, models are no better than theories, concepts and propositions. (And this is why I have been including models with theoretical structures). Portraying theories as sets of models does not solve the question of how structures get empirical content. I do not mean to imply that focusing on models does not help; I am merely pointing out that switching focus from propositions to models does not solve the problem on its own.

Here are five solutions to the problem of how we coordinate structures (including models) with their target systems. They are inspired by 1) empiricism, 2) conventionalism, 3) Kantianism, 4) Quinean holism and 5) a contextual model-based position.
The first solution is to see theory as coordinated with empirical content because theory is always originally built up from experience, beginning with some very basic experiences. Arithmetical truths are grounded in our experience with counting, and other scientific terms are grounded in equally unproblematic experiences such as moving through space (distance) and experiencing change (time). This is a very old empiricist program, endorsed by Mill and some of the logical positivists. Ernst Cassirer emphasized its main flaw, which Reichenbach missed. You cannot begin with “raw” experience, because there is no such thing (see Cassirer 1910). As soon as we begin talking of moving through space, we assume already the concept of space. One orange plus one orange entails the possession of the concept ONE, itself a difficult concept. And so on.

The second approach is conventionalism, according to which we can stipulate what a theoretical structure corresponds to, by fiat. This allows scientists to “get on with it.” Reichenbach saw that this would not do, however, because in such a stipulation you assign one thing to another (van Fraassen 2008, 120). We know a lot about the one thing; it is a well-defined theoretical structure composed of concepts, representation relations, mathematical formalisms, models, etc. But the other cannot be described except by already presupposing the theory (or some theory). We cannot relate a theoretical entity to a physical entity without knowing what the physical entity is to be related.

The third option is to suppose that some principles of coordination are conditions of the possibility of scientific practice including measurement. This is the Kantian answer. There are at least two ways to develop this strategy. One is Michael Friedman’s, according to which some propositional stipulations must be taken for granted in order for a theory to have empirical consequences at all (Friedman 2001). These propositional stipulations include basic axioms, mathematical tools and philosophical assumptions. While it might be true that such stipulations will be necessary for theoretical structures to have empirical content, it does not explain how those stipulations will have content. If they do not have content of their own, then it is hard to believe that they can provide content, unless by stipulation, which we have already rejected. The other way to be a Kantian about coordination is to take a certain set of cognitive abilities as necessary to make experience possible (see Buzzoni 2008, 2010). For example, Buzzoni argues that our ability to reason counterfactually is what makes experience meaningful. This answer will be considered in more detail below. I think this is a good answer, but again it is insufficient on its
own. Buzzoni, for instance, must combine his Kantianism with operationalism to provide the empirical content for theoretical structures.

A fourth solution is Quine’s (1951): we can dodge the problem of coordination by denying that scientists need to bridge their structures to other structures or to experience because no strict distinction can be drawn between form and content (or theory and experience) with respect to meaning or truth. In other words, each node in our web of knowledge depends for its meaning and justification on some empirical and some theoretical matters. Heisenberg’s uncertainty principle, for example, is already empirically meaningful because each of the concepts that compose it already contain information about the world by virtue of their existing as part of our interconnected web of knowledge.

However, even assuming that everything is always already partially empirical, this would not solve that the problem of coordination. This is because the problem is not about how we fill empty formalisms with raw content. If it were, we could simply deny that there are empty formalisms. But as long as there is the possibility that applying what we have developed in one part of our system of knowledge to what appears or should be expected to appear in another can be difficult and desirable, then the problem of coordination will persist. In other words, we do not need a strict distinction between theoretical and empirical matters for stronger connections between them to be desirable.

To be more specific, let us consider the Quinean response more closely. We can consider Quine’s web methodologically, epistemologically or semantically. The methodological interpretation considers how we revise our web when parts of it turn out to contradict one another: we must revise all of it (or much of it). This interpretation does not conflict with the need for coordination: revision can and does prompt new issues of coordination. The epistemic interpretation concerns the source of our justification: knowledge is never gained in isolation. But this does not mean that there is nothing more to learn about the relationships between pieces of knowledge. I may know what happens in Orwell’s Animal Farm, but I might not know how it is a criticism of Stalin-era Soviet Union policies. I may learn the basics of Spacetime, but not be able to say how it will act near a black hole. Finally, if we understand the web semantically, then the meaning of each term is related to the meanings of others. But semantic relatedness does not
guarantee semantic determinability. Some concepts (like TRIANGLE) might be semantically determinable given their relations to other concepts (like THREE, SIDE, CLOSED, SHAPE, etc.). But others concepts (like JUSTICE), do not work this way. While these more complicated concepts are related to other concepts, those relations do not express the entire semantic content. We could try to rephrase JUSTICE in terms of RIGHTS, FAIRNESS, AUTHORITY, etc., or as a set of relations between institutions and individuals in an ideal society, or a set of purported modern and historical instances. No attempt at the first strategy has been successful, and even if it were, it would only deflect the question. The other strategies fail because they ignore the concept’s normative significance. Philosophers are more interested in what should count as JUST, rather than what people currently count as JUST, or might count as JUST. Connection to experience is not specified by saying that our concepts are connected in a web. Quine says that the web impinges on experience at the edges of the web, but we should ask what it means for webs to impinge.

The final family of solutions to the problem can be found in the literature on modelling. There are two main solution-types. The first concerns representation relations. Some claim that models represent their target systems by isomorphism (van Fraassen 1980) or partial isomorphism (Da Costa and French 2003), or by respects and degrees of similarity (Giere 1988), which are more flexible than isomorphism. Some respects of similarity include shape (for example, in physical scale models), material (as in laboratory models), and dynamical structure (order of events). The basic idea behind all these accounts is to see the connection between model and target facilitated not by propositional bridging principles as in Carnap, but mapping relations like isomorphism or similarity.

The second solution is to say that models do not need to be connected to their target systems by some action or relation, because in some sense models already possess the important features of their target systems. The model preserves the modal structure of the system, or its material features, for example. These preserved features are what is left after the model has distorted the system via abstraction and idealization. Abstraction removes features of a system that are deemed irrelevant or confounding. Idealization positively distorts features of the target system, for example, by making a system frictionless or (in)finite, etc. But some information sneaks past the abstraction and idealization process, and because of this, coordination is not necessary.
However, neither of these solutions is successful on its own. Take the first. Model representation by (partial) isomorphism or similarity involves connections being made. Which connections we make is not something that could be mandated by the model, the system, the background theory, general scientific norms, or the combination of any of these. And we would not want it any other way, because not having a complete account of all the model-target relations is scientifically advantageous. Hesse (1966) has argued that models will have positive, negative and neutral analogies to their target systems. A positive analogy is one that we know to obtain, and it is on the basis of analogies like these that we construct the model to begin with. For example, the laws that govern the changes in the position and momentum of colliding billiard balls provide a positive analogy for the changes in the position and momentum of particles in the kinetic theory of gas. Negative analogies are analogies that we know not to hold. For example, billiard balls are striped or solid in colour, particles are not. We should avoid making inferences about a target system that draw upon the negative analogies of the model. However there are also neutral analogies. These are analogies which are not obviously positive or negative. It is often because of neutral analogies that we learn something new and surprising. For example, the Michelson-Morley experiment showed that an important analogy between light waves and water waves was actually a negative one. Namely, the analogy does not hold with respect to the need for media. Water is necessary for water waves, but a luminiferous ether is not necessary for light waves. The fact that we do not know whether or to what some features of a model refer is a source of progress. But it is also the source of the problem of coordination (for models). Again, I am not arguing that this problem is insurmountable; we solve the problem all the time. However it is a problem.

The second model-based strategy is to argue that models retain aspects of the target system. Notice that as soon as a natural system is made abstract, it departs from reality. This on its own is not a problem: at this stage there is nothing to coordinate that is not already coordinated. An abstract predator-prey model has all the same empirical content as the real relationship it models, only less, because it does not include influences from the rest of the ecosystem. However, the moment that model is used to generate inferences, the question of coordination arises. What happens when we increase the predator population in the model is not necessarily what would happen in the target system (even assuming some of the real system’s information is retained in the model).
Also, in cases where we are modelling complex target systems, some coordination will always be necessary to tell us what the features of the post-manipulation model refer to, and perhaps more importantly, to relate the mechanism that produces conclusions in the model (deduction, computer simulation, etc.) to the mechanism of change in the target system. This is necessary to justify conclusions made about the target system. We cannot say what makes a good model by referring only to the internal features of the model; we must also consider its relationship with the target, and how it differs.

A similar situation obtains for idealized models, except in this case, the differences between the model and its target system are far greater from the start. An ideal pendulum has different empirical content from a real pendulum because it changes the amount of friction present, the distribution of the mass of the bob, and the rigidity of whatever connects the bob to its origin. The swinging ideal pendulum (and the inferences it that licenses) are not the same for the swinging real pendulum. And this is a simple case; the sort of idealizations made elsewhere in science are far muddier. Also we must keep in mind that any application of mathematics will involve idealization. Whenever we assign a number to something, we assume an infinitely precise measurement which is all but certainly inaccurate. When I say that I am 183 centimetres tall, I am saying that I am 183.000000000000000000… centimetres tall, which has a probability that approaches zero (Duhem 1954, 132-143).

One last comment on models: a very interesting claim in the philosophy of science is that models are independent of theory. Margaret Morrison has called models “autonomous agents” (1998). If this is correct, independence of a model from theory might make coordination necessary both between model and world and model and theory. This might have been expected, since I focus on the relation not only between theoretical structures and experience, but also with other structures. Still, a few words are necessary.

Models can be constructed to help a theory yield precise predictions. But the theory does not tell you how to create or use such a model (Cartwright 1999, 179-210). “Model building is an art and not a mechanical procedure” (Hartmann and Frigg 2012). This makes model construction at least somewhat independent of theory, which creates room for coordination. The model need not capture every feature of the parent theory that it exemplifies, and there are a wide variety of ways...
that a model could be built to exemplify the very same theoretical features. Because of this, the model does not coordinate itself; we coordinate it as we build and use it.

I conclude that contemporary and historical solutions the problem of coordination are at best incomplete. Very recently, combinations of the above solution-types have emerged in the literature on measurement. To begin with, here is how the above solution-types look when phrased in terms of measurement. 1) Theory is coordinated from the bottom up through measurements of quantities like space, time, etc. 2) Theory is coordinated when we stipulate measurement indices, as when we set 100°C as the boiling point of water. 3) Theory is coordinated by principles that are already always in place to make science possible, which concern how to access the world via measurement actions. 4) Theory does not need to be coordinated because it is situated in a web, the borders of which impinge on the world via measurement. And 5) measurement models serve as mediating entities between theory and world via representation relations.

Here is how recent work by Bas van Fraassen, Hasok Chang and Eran Tal have combined attractive aspects of these answers.

van Fraassen understands the model-based solution as taking one of two forms: the ahistorical “view from above” and the historical “view from inside.” The first determines what a theory was measuring all along given our current standpoint (or some future standpoint). The second determines what a contemporary theory measures given the contingent historical practices of contemporary scientists. The ahistorical view from above would claim that when Galileo created thermometer-like instruments, he found a way to measure with some accuracy what we think of as temperature and pressure. If we take the historical view from the inside, however, we consider early thermometry practices from the perspective of Galileo’s time and claim that early thermometers were sensitive to changes in heat and altitude. Different liquids and gasses were tested to see which corresponded more reliably with which changes in heat, and different types of apparatus were designed to contain them. These apparatuses also had different ranges of functionality. Each apparatus was used to test the others, and eventually some stable regularities were perceived, including the relation $PV = nT$. From the inside we frame the developments in terms of problems, concepts and methods contemporary to the period.
Most of the time, the ahistorical view from above and the historical view from within will select different targets of measurement for the same processes. But van Fraassen points out that in both cases, coordination relies on a stable combination of theoretical assumptions and measurement practices from which to judge the referents of measurable quantities. The difference is that the ahistorical perspective considers its target from a different historical context defined by different practices and assumptions, while the historical perspective uses the same context.

Chang (2004) argues that coordination takes place through contextual processes that mutually and simultaneously define measurement practices and theoretical entities. What we see are successive iterations of measurement apparatuses and theoretical predictions, which are re-coordinated and updated at each stage. “The challenge for these writers [van Fraassen and Chang] is not to find a vantage point from which coordination is deemed rational a priori, but to trace the inferential and material apparatuses responsible for the mutual refinement of theory and measurement in any specific case. Hence they reject the traditional question: ‘what is the general solution to the problem of coordination?’ in favour of historically situated, local investigations” (Tal 2012, 10).

Tal follows van Fraassen and Chang with respect to measurement coordination, but he wants to provide a more general account that is not limited to case studies. He argues that “coordination is…a process rather than a static definition” (2012, 11). Tal argues that measurement coordination is mediated by models, which are abstract representations of local phenomena (2012, 17). Specifically, measurement happens when we “estimate the value of a parameter in an idealized model of a physical process” (2012, iii). Given some measurement procedure, we infer to the value of a parameter by the use of the theoretical and statistical assumptions in the measurement model. This model is of an idealized version of the procedure. If it were not, we could not justify the inference to the parameter. That is, idealization is a necessary precondition for the possibility of justifying the measurement as a measurement of the given parameter. In other words, idealizations are necessary for solving the problem of coordination in all cases of measurement.

This contextual model-based solution to the problem of coordination is attractive because it makes room for scientists to stipulate their referents (as in the conventionalist solution), but only
tentatively, in a dynamic and iterative practice of model-building that relies for its possibility on idealization (as in the Kantian solution). It takes seriously the methodological, semantic and epistemic web of Quine, and it makes use of mediating models that lie between theory and target.

However, Tal’s solution to the problem of coordination via measurement (to focus on idealization) is still not enough to solve the general problem of coordination between theoretical structures and their semantic content. Let us assume with Tal that in every case of coordination for a theoretical structure there will always be idealized assumptions to facilitate that coordination. The question, “How do we coordinate such models?” is still pertinent. Surely Tal and Chang are correct that scientists try out different idealizing assumptions and update their models as they update measurement apparatuses. But neither this practice nor the existence of idealization will be the last word in coordination. As I argued above, idealizations are often the reason we have to coordinate in the first place.

The question of how scientists create and use idealized models to make contact with the world is one that can only be answered with reference to the abilities and actions of agents. The agent is the most important part of the modelling process, and the ability of the agent to idealize and work with models is the most important capacity of the agent as far as coordination is concerned. This brings us back to the studies with which I began this chapter. I showed that these studies agree about what is important for understanding, namely, the active creation and manipulation of the imaginary scenario by the agent. There is no such thing as coordination in the abstract. Theoretical structures are not coordinated to systems in a general sense, in the passive voice. Coordination is something we have to do.

Unfortunately the studies in this chapter do not tell us which cognitive abilities make coordination via thought experiment possible or successful. The next chapter therefore introduces an account meant to fill this gap. At the center is the cognitive ability of the agent to imagine.

3 Conclusion

Studies in social and cognitive science agree that thought experiments bridge conceptual/theoretical knowledge to kinaesthetic or kinematic knowledge. The thought
experiments cited in these studies were used to explore or interpret solutions to problems, communicate ideas, or model scenarios, more than they were to provide solutions. All the above studies support the claim that by increasing intelligibility and fruitfulness, we can increase our understanding of a theoretical structure. I also suggested that the first of these is a prerequisite for the second: if we do not make sense of a theoretical structure, we cannot use it.

These considerations raised the following question: how epistemically necessary is understanding in science? I answered in the second half of this chapter that, if scientists want to claim they possess knowledge about the real world, they must have a way to solve the problem of coordination when it arises. Since the problem of coordination results from a lack of understanding, and solving it is necessary for justifying science’s grip on reality, understanding is necessary for scientific knowledge. I also showed that the problem of coordination does not yet admit of a general solution in philosophy down to the level of the agent, which is what is necessary. Since it seems that thought experiments help us to solve the problem of coordination, and thought experiments are performed by agents, an agent-focused account of scientific thought experiments is the aim of the next chapter.
Chapter 7
Imagination and Understanding

Among so many unforeseen developments in cognitive science and neuroscience, surely one of the most surprising concerns the recent surge of interest in imagination as a subject for scientific research and speculation.

(Richardson 2011, 663)

We saw in previous chapters that scientists including Mach, Reichenbach, Einstein, and Heisenberg recognized the necessity of coordinating theoretical structures with empirical content. I presented evidence that scientists use thought experiments to achieve this goal (Chapter 5), and I hinted that perhaps the key to achieving it lies in the imaginative use of idealized models (Chapter 6). It is time to draw the threads together. If Tal is right that idealized models are the key to connecting measuring devices and their targets, and I am right that the agent should be at the center of the solution to the problem of coordination, then we must begin by asking how agents manipulate idealized models in their minds.

Tal’s discussion suggests two ways a model can help us navigate the connection between a theoretical structure and experience. The first is conventional. For example, a student might wonder about the empirical significance of $E = mc^2$ and we could directly provide the definitions of $E$, $m$, $c$, squaring and $=$. The second is constructive. For example, we could have the student consider what happens to mass as it approaches the speed of light. As you increase the kinetic energy of an object, it speeds up (say, according to $E_k = \frac{1}{2} mv^2$). But at the speed of light, we cannot increase the speed of the object any more. So what happens when we continue to add energy? Either the mass or the velocity must increase, and it will not be the velocity. Therefore, mass must be increased by the addition of energy. Hopefully this mental exercise makes the mass-energy equivalence intuitive. This is one way the empirical significance of a theoretical structure can emerge organically through thought processes. The suggestion in the last chapter was to see models as a middle ground that allow for both convention and construction. And
certainly either method of increasing the empirical significance of $E = mc^2$ could in principle succeed.

We should recognize before moving forward, however, that the constructivist option is more fundamental than the conventionalist one. If a student understands $E = mc^2$ through reference to $E, m, c$, squaring and $=$, the question remains how that student came to understand those variables and relations. It cannot be stipulation all the way down. This is not a foundationalist claim; meaning does not have to originate from some bottom-level meaning-making experiences. It can be created at any level (see Sayward 2000 for a discussion of conventionalism without foundationalism). For example, when physicists utilized civilian help to interpret bubble chamber photographs, the civilians were taught to recognize particle trails without necessarily knowing what the particles were (Galison 1997, 553–688). Being aware of properties of the target of a theoretical structure, relating that structure to other concepts and experiences, and having the ability to use the structure in a way that gains the approval of others who already understand it, are exhibitions of new understanding. Regardless of what level we create meaning on, it will be the constructive creation of meaning that answers the general problem of coordination, not convention. I will focus therefore on the historical-experimental (constructivist) solution to the problem of coordination as the more fundamental solution, since the conventionalist response relies on it.

How is the empirical content of a theoretical structure increased for an agent through thought experiments conceived of as mental manipulations of models? I begin by arguing that the accounts of Tal, Nersessian and Miščević, while perhaps successful for their purposes, do not explain this use of the imagination in science. After a brief look at some results in neuroscience concerning “bottom-up” and “top-down” reasoning, I present my account, which explains transcendentally the role of the imagination in increasing understanding. I close by considering some consequences of this account for the literature on scientific representation and understanding.

1 Mental Models

There are several ways that mental models could account for the understanding produced by thought experiments. One would be analogous to the role played by the measurement models.
considered by Tal. These models make possible the connection of objective measurement outcomes with local and idiosyncratic instrument indications (Tal 2012, 178). Thought experiments can be seen as performing a similar role when they connect abstract theoretical structures to the local and idiosyncratic experiences of the agent performing the thought experiment. Tal might account for this by reference to the way models employ idealized assumptions. This will not do, however, as an explanation for thought experiments that increase understanding. Suppose a student wanted to understand the principle of equivalence. We present the student with Einstein’s elevator thought experiment. Suppose the student still does not understand. What do we do? It is likely that we would have the student perform the thought experiment again, more slowly, asking them to put more effort into imagining the events described in detail. Or we might create a simpler thought experiment to introduce one of the more difficult aspects. For example, we might consider what it feels like to be in an elevator on Earth that is moving upward, before we consider being in an accelerated elevator in a gravitational vacuum. In either case, it does not seem that it is the idealizing assumptions themselves that are responsible for the connection to be established, but rather something the agent does with them. We can point explicitly to the idealizations—the perfect acceleration and perfect gravitational vacuum for the one scientist, the perfectly uniform gravitational field for the other, etc.—but without putting him or herself in the shoes of both scientists, the thought experiment will not work.

Secondly, the thought experiments I am considering in this thesis succeed only because they keep one foot planted firmly on solid ground. Their idealizations and abstractions take us into fantasy, but they are constructed so we do not lose track of the real world. For example, the equivalence principle is completely general and abstract, and Einstein’s elevator adds content to it by providing a local case that while partially idealized, illuminates how such a principle might be instantiated in the real world. Darwin’s eye, Heisenberg’s microscope and others do the same. The thought experiments add content to our understanding of theoretical structures by creating local possible experiences that instantiate them. It is possible to portray the imagined scenario as an idealized and abstracted version of a real experience, but that is not what is crucial. Einstein’s elevator provides an example of the equivalence principle that helps us understand and apply it, it does not invite us to consider an abstract and idealized model of someone performing experiments in elevators. Maxwell’s demon exists to enlighten us about statistical
thermodynamic fluctuations, not about what a tiny person with strange eyes and a frictionless door does in a room full of particles.

Additionally, thought experiments do not necessarily provide more understanding when they deviate further from reality by idealization or abstraction. Galileo’s falling bodies thought experiment and boat-cabin relativity thought experiment only deviate slightly from reality (removing air resistance and heavy sea waves, respectively). The first would not be better if instead of cannon balls and musket balls we considered point masses, and likewise for the second if we imagined carrying out idealized experiments in the boat cabin, like calculating the motion of a swinging ideal pendulum whose anchor point had a constant velocity or no velocity with respect to some other fixed point. In fact I think these alterations would make the thought experiments less successful.

Finally, the idea that thought experiments coordinate theoretical structures by the use of idealized models is not satisfactory because in order to make use of such models we will need more idealized models, ad infinitum. To say that we provide content for abstract structures by the use of other abstract structures only deflects the question of how we provide content for abstract structures. Whatever we take models to be, as long as they are abstract structures of any kind, I do not think that they can perform the function I attribute to them in this thesis. We can connect abstract structures to more abstract structures forever without getting any grip on what it all means. This is the reason cognitive scientist and philosopher Alva Noë has recently written that

Clocks may keep time, but they don’t know what time it is. And strictly speaking, it is we who use them to tell time. But the same is true of Watson, the IBM supercomputer that supposedly played Jeopardy! and dominated the human competition. Watson answered no questions. It participated in no competition. It didn’t do anything. All the doing was on our side. We played Jeopardy! with Watson. We used ‘it’ the way we use clocks.

Noë contrasts clocks and computers with amoebas, which “are smarter than clocks and supercomputers. For they possess the rudimentary beginnings of that driven, active, compelling engagement that we call life and that we call mind. Machines don’t have information. We process information with them. But the amoeba does have information—it gathers it, it manufactures it” (Noë 2014).
In the same way, models are tools that we use. They are like programs that we can run in our minds. We can fill in these programs with more information, more assumptions, more representations, but these will never make contact with their targets without what Noë calls “that driven, active, compelling engagement that we call life and that we call mind.”

Assuming we want to keep characterizing thought experiments as mental models, where do we locate their “compelling engagement”? The answer must be in the performance of building and manipulating models. We must create the model for ourselves every time we use it, and we must connect it with experience. There is a sense in which empirical science has already provided the vocabulary to discuss what goes on when we actively create and run mental models. For instance, we can refer to changing patterns of activation in neural networks and the establishment or atrophy of neural pathways. Miščević and Nersessian pursue this option as a way of explaining the cognitive efficacy of thought experiments in science.

Miščević (2007) argues that the power of the imagination results from the manipulation of mental models, which help us to “mobilize unarticulated knowledge.” This is how we learn from thought experiments. The conclusions that result from the exercise of this capacity are justified by the fact that such a capacity has evolved as a useful predictive tool with its roots in normal perception. Nersessian agrees, stating that “the perceptual system plays a significant role in imaginative thinking,” which “makes sense from an evolutionary perspective” (2007, 136). While Nersessian does not claim that all the content that is manipulated in mental models is perceptual or imagistic (142), she does “contend that a wide range of empirical evidence shows perceptual content is retained in all kinds of mental representations” (139). What grounds the epistemic use of thought experiments for Miščević and Nersessian is therefore experience itself, combined with reliable cognitive and sensory faculties that evolved to function in non-imaginative cases, but remain relatively reliable in imaginative ones.

The idea that we manufacture complex ideas from sensory experience has a lot of empirical support (for example, see Prinz 2002). Nevertheless, I think there is something about the use of the imagination in producing scientific understanding that is left out when we consider thought experiments as ordinary manipulations of existing sensory experience. We might be right to trust knowledge claims based on what happens in imaginary scenarios because they are formed by
cognitive faculties that are reliable. But in producing understanding we attempt to create empirical content that is intelligible and fruitful, and this is sometimes best achieved by letting go of empirical accuracy. There is no observation I can make that would explain to me the empirical significance of Einstein’s principle of equivalence. And increasing the accuracy of all my representations would not help me. Einstein’s elevator succeeds because the imagination takes us away from what we have experienced and tries out new links between existing concepts, experiences and models. I said above that thought experiments keep one foot in reality by providing possible local experiences to add specificity to general theoretical structures. But they also help us leave it behind. This is not contradictory; we have to transcend our existing experiences to see how a principle can connect to experience generally, but we also need specific examples to show us how this can be done in a single case which may be inductively applied elsewhere.

Since this use of the imagination is very different from the one that generates new knowledge, we must search for a different explanation. The key is recognizing that such an explanation will not be either a priori or a posteriori. We cannot explain our ability to connect concepts to their instances, or models to their targets, or theories to their models, without being able to connect the notion of connection to each of these.

2 A Transcendental Account of Understanding-Oriented Thought Experiments

In this section I will argue first that since the imagination can give content to some acts of basic perception and cognition, thought experiments can give content to theoretical structures of science. Second, I will argue that none of the usual ways of knowing provide understanding in the sense I am interested in, but the imagination does. This makes imagination necessary for any act of experimental or theoretical understanding.

Studies in cognitive science support the notion that imagining is important for cognizing. For example, the now classic Miller and Cohen (2001) discusses top-down processing, which has become “central in modern neuroscience” (Burchard 2011, 69). Top-down processing occurs when we first categorize or cognize things in broad strokes, and work through the details later. These details can include the specific features of the general object-type. First we see a rabbit,
and only afterwards do we see its long ears and bushy tail. This means that the higher centers of our brain actually help to determine what we experience. When top-down processing is operative, higher centers in the prefrontal cortex of the brain track and modify what happens in lower centers. When something new or difficult to identify is presented to a subject, top-down processing begins before recognition of the object is accomplished. According to Miller and Cohen (2001), this occurs when the prefrontal cortex provides bias signals to lower brain structures. These bias signals guide the flow of neural activity along certain pathways. In other words, when we see something, parts of our brain normally associated with conscious thought are already involved in determining what we perceive (see also Buschman and Miller 2007).

Here is some of the evidence for this phenomenon that is relevant to imagination’s role in visual cognition. First, there are many documented cases of patients who have lost their visual cortex due to injury, or who were born blind, and yet can imagine things vividly (Arditi, Holtzman and Kosslyn 1988; Kerr 1983; Marmor and Zaback 1976; Nersessian 2007, 137). Therefore perceptive imagination does not only depend on the visual cortex. Second, there are patients who have damaged parts of their neocortex that are normally not associated with vision, and yet cannot see conceptualized objects (such as birds). They can only see lines, shapes and patches of colour (Thagard 2010a, 70). This shows that the neocortex, which is responsible for higher forms of cognition, plays a crucial role in perception. Finally, if we approach an ambiguous figure with a certain interpretation in mind, this often determines what we see when we look at it. All of this is important evidence that the imagination can play a role in determining the content of perceptual experience. Of course in a healthy individual, both directions of interaction mesh together; bottom-up and top-down cognition are not step-wise, or even separable. Thagard argues that “because brains perform inferences using parallel activity of millions of neurons, perception can elegantly integrate both bottom-up and top-down information” (2010a, 71). And “perception involves simultaneous parallel processing that combines top-down knowledge with bottom-up perceptual input” (2010a, 100).

Top-down processing shows that there is an interesting connection between higher cognitive faculties and the most basic content of thought. This goes not just for perception, but other reasoning tasks including categorization. To explain the role of higher cognitive faculties in categorization, Mark Johnson introduces the notion of “schemata.” Johnson claims that we
“connect up” (1987, 152) abstract mental structures (like concepts) with the contents of our sense perception using schemata, which for him are “nonpropositional structures of imagination” (19). Johnson maintains that “even our most simple encounters with objects, such as the perception of a cup, involve schemata that make it possible for us to recognize different kinds of things and events as being different kinds” (20). Johnson’s schemata have been very influential in cognitive science, and after the idea resurfaced in Lakoff and Johnson (1999), it spawned a subfield of research. The basic idea is that schemata, which are structures of the imagination, give content to our concepts, and thereby help to structure thought. How do they work? By exercising the imagination, which for Johnson is “our capacity to organize mental representations (especially percepts, images, and image schemata) into meaningful, coherent unities” (1987, 140).

Nigel Thomas understands schemata as data structures in the brain that make possible perceptual experience of the world (Thomas 1999). This is slightly different than the Johnson-Lakoff school’s characterization, but Thomas admits that the views are compatible, and again the imagination plays a central role in both perception and reasoning. Thomas argues that schemata are not things that we experience, although they are necessary for experience in general.

I very much approve of this emphasis on the imagination as a capacity that structures perception and categorization. However I do not think that philosophically speaking the schema should be considered as “structures of the imagination” or “data structures.” First, it is not at all clear where these structures would reside. Are they mental or neural? Can we point to them? Second, it is not clear what causal mechanism the schema has that allows it to structure things. Without a more detailed picture of this process, Johnson’s account merely says that we categorize by structuring our percepts into coherent wholes, which we do through the faculty of imagination, which is the faculty that allows us to structure percepts into coherent wholes. Third, suppose that a concept is a structure (like a proposition or a model), which has relations to other structures, and also has a (set of) referent(s). The problem of coordination asks how structures like these are given empirical content. If the schema is another structure, we are faced again with an infinite regress. In other words, while a structure can structure structures, this does not solve the problem of coordination without something that gives (some of) those structures their empirical content. For this reason, I will leave Johnson and Lakoff’s schema behind, though I fully accept their empirical work concerning the importance of imagination for categorization.
To summarize, the imagination does occasionally appear to help determine the meaning of some thoughts and perceptions. We do not always need to imagine a duck to see one, but we might need to imagine an electron to see electron trails in a bubble chamber. This brings me to my main point. Here are four reasons to accept the idea that thought experiments can provide content for theoretical structures of science (3 and 4 are inspired by Buzzoni 2008):

1) The imagination can help to provide content for simple perception and cognition, so it should also be possible that it could do the same for the more complex theoretical structures of science. I can think of no relevant difference between the features of perception, experience and cognition, on the one hand, and the concepts, models, theories, and principles of science on the other that would enable us to say that the imagination can provide content for one and not the other.

2) What seemed hopeful at the end of the last chapter was the necessity of idealization for providing empirical content to theoretical structures. But I rejected idealization as a candidate when considered on its own, and argued that what mattered was the ability of the agent to make and use idealized models. What accounts for this capacity? Idealization (and abstraction) depend principally on our ability to see the world as being different than it is. In other words, idealization and abstraction depend on our ability to imagine.

3) Other candidates fail as possible bridges between theoretical structures and experience, knowledge and abilities. Mathematical derivation and logical inference will not ground empirical content. A computer simulation must have its semantic content programmed in advance. The performance of an experiment on its own will not provide empirical content either, because the action of performing an experiment is simply people doing things and taking notes. Without conceiving of the experiment as testing a theoretical claim, all of the actions involved will have no theoretical relevance, and it will not provide semantic content to the theoretical structures involved. What allows us to use any of these tools of science in a way that bridges theory and experience is the imagination.

4) If Chapter 3 and 4 are correct, then inference-making and computer simulation both receive much of their justification from the fact that it is through trial and error that we first recognize what works in these methods. This means that the most basic epistemic
tools of science are experiment and observation. Experiments vary variables in carefully contrived ways. Varying a variable is only possible given our recognition that certain events have a “natural” tendency that we purposely disrupt. Seeing the possibility to disrupt these tendencies relies on seeing that they can be interrupted, or in other words, seeing that the world can go in different ways. And this relies on the imagination. The imagination is therefore what makes experiment possible in general. Concerning observation, we already saw above that the imagination is an important part of both perception and categorization. Therefore imagination is at the heart of the most fundamental tools of knowing.

I conclude that imagination is a *sine qua non* of science. Without it, human scientific understanding would not be possible.

Now, nothing I have said above necessitates that the output of any scientific experiment, inference, model, etc., will be *justified*, simply by virtue of having gaining its empirical content through imagination. Does imagination merely make science possible, or does it also provide reason to trust it? How do we distinguish between helpful and hurtful uses of the imagination to increase empirical content?

Attempting to answer questions like these brings us to the transcendental nature of the imagination. If theoretical structures require the imagination for us to understand them, then the use of the imagination in science cannot fully be investigated by either cognitive science or philosophy, since it is always already presupposed by both. We cannot explain (without using the imagination) how the imagination connects theoretical structures to content because we need the imagination to understand what we mean by THEORETICAL STRUCTURES, IMAGINATION and CONTENT, which are themselves theoretical structures. We cannot understand how imagination provides understanding without imagination providing understanding.

Let me repeat this point using an example from Chapter 3. There I discussed Lewis Carrol’s paper “What the Tortoise said to Achilles.” Carrol argued that we cannot justify modus ponens (our most important of inference rules) without already knowing how to use modus ponens, and accepting it as a valid inference. What is *really* the basis of our justification of modus ponens is this fact: that we must accept it even to consider that it might be false. Any argument against
modus ponens will assume the truth of the premises and reason to the truth of the conclusion, or in other words, instantiate modus ponens. Even if we only assume modus ponens for reductio, the final argument against modus ponens will read as follows: if we can reduce modus ponens to a contradiction, modus ponens is invalid. We reduced modus ponens to a contradiction. Therefore, modus ponens is invalid. But this argument is nothing other than an instance of modus ponens. The same holds of the imagination. If we accept that the imagination helps to determine the empirical significance of theoretical structures, then we cannot coherently consider that it could to do this, since we would need the imagination to help us understand the terms in that argument.

If we have no choice but to allow for the role of the imagination in grounding empirical content for theoretical structures, then we should ask how the imagination is used to increase empirical content. Given the case studies in Chapter 6, I think it is likely that understanding-oriented thought experiments increase empirical content by inspiring us to make and try out various actions of connection. While human imagination can automatically and subconsciously connect theoretical structures to empirical content (in simple cases of perception and categorization), there are cases in science when our theoretical structures cannot be connected in this way. The theoretical structures might be too novel, too unconnected to empirical subject matter, or they might invoke a combination of notions that we have a difficult time connecting, or ask us to disconnect two notions that we have a hard time disconnecting. Consciously-built thought experiments are then introduced to facilitate this action by forcing us to focus on the narration of an imaginary scenario, which capitalizes on our capacity to reason using fictions.

There are two important points to keep in mind. First, a thought experiment is not its presentation in a book or article; it is the performance of following a narrative, leading to the creation of an imaginary scenario that is manipulated somehow or other, resulting (we hope) in the achievement of some epistemic good. Second, while cognitive science can explain the performative aspect of a thought experiment in terms of the electrical activity of networks of neurons, this is not what is responsible for the understanding created thereby. It is certainly possible that thought experiments are multiply realizable. I would not want to dismiss the possibility that intelligent agents whose centers of intelligence are otherwise constituted or
organized could connect their experiences and theoretical structures, perhaps by thought experiment.

To summarize: active use of the imagination (something akin to top-down reasoning) must solve the problem of coordination by increasing the semantic content of theoretical structures. The power of the imagination to do this can be been grounded by a transcendental argument. We cannot use empirical research (cognitive science) or a priori reasoning (philosophy) to deny that the imagination increases empirical content, because in both cases we will rely on the imagination to understand the terms, processes and results of our investigations. This does not imply that given an acceptance of the role of the imagination, we cannot use philosophy or cognitive science to investigate it. In what follows I will provide some philosophical analysis. In future work, I will turn to cognitive science.

There are three philosophical issues that will be addressed in the rest of this chapter: 1) the nature of empirical content, and how it relates to meaning and reference, 2) the nature of understanding, and how it relates to knowledge and objectivity, and 3) how using the imagination to provide understanding can go wrong.

3 Empirical Content Revisited

I begin by comparing the interpretation of theory to the everyday interpretation of language. A short argument shows that there is only a difference of degree, not of kind. This will allow me to use resources in the philosophy of language in the discussion that follows.

In the same way that idealizing assumptions are necessary for the creation of mental models, but are not epistemologically responsible for the power of thought experiments to increase understanding, intuitions about what a speaker means are necessary for linguistic interpretation, but are not responsible for its success. When we encounter a linguistic item (utterance, term, etc.) that we want to understand, we begin by making imaginative guesses. These are constrained by background knowledge and a principle of charity. We test these guesses by trying them out, refining them, and testing them again until we are satisfied. If our intuitions completely were limited to what we meant by our words and what we believed to be true about the world, this would prevent us from successfully interpreting the utterances of people who have different
beliefs or assign different meanings to their words. Given that we can interpret other people, we must be able to leave behind the meanings of our own words, and imagine what it would be like if they were assigned different meanings. This brings us again to the power of the imagination. We understand others by finding out what they mean, which we do by first assuming that they mean what we do by their words. If this does not work, we create and test possible meaning ascriptions for our speaker, which we accept or reject in an iterative way that is updated with new experiences. And this is precisely the same process that goes into understanding new theoretical structures or developments in science: we assume our usual interpretations of concepts like TIME, SPACE, INFORMATION, FORCE, etc., and only when these do not work, we begin trying out alternate assignments of empirical content for those terms or events. Our interpretation of a speaker’s words is therefore just as experimental as our scientific interpretation of theoretical structures and experiments/observations. And all of these depend on the imagination.

Now, is there a distinction between empirical content (as I have been using it in this thesis), and meaning (as it is used in the philosophy of language)? There does not have to be, although such a distinction might be helpful. Empirical content as outlined in Chapter 1 is an umbrella term. We can think of it as the pairing or mapping of terms to extra-mental referents, the empirical psychological “sense” of something, or a list of what would need to be the case in the extra-mental world for a proposition to be true. The term “meaning” has been used to denote each of these.

Furthermore, both meaning and empirical content change drastically when applied to terms, propositions, theories, actions and events. The meaning of a term can be an individual’s sense of that term (my sense of HORSE includes an earthy smell and hair that feels less soft that I expected), or a function that relates that term to referents (causally or otherwise), or the extension of that term in one or two dimensions of possible worlds (see Chalmers 2006). It could also be a prototype, exemplar or mini-theory of the type of thing denoted or connoted by the term (see Margolis and Laurence 1999, Prinz 2002). The meaning of a proposition could be something like a picture, or the bearer of a truth-value, or a thought. The meanings of theories (as sets of propositions or models), actions and events are even broader. This is also the case for empirical content. The empirical content of a term is not the same sort of thing as the empirical content of a
proposition, theory, action or event. But whether we portray the content as relational or psychological or representational, it will be empirical, it will be possessed by an agent, and it will connect the term/proposition/theory/action/event to something else, like experience, other terms/propositions, abilities, values or emotions.

Despite these similarities, I have tried to keep empirical content separate from meaning in this dissertation because meaning is commonly associated with (Fregean) psychological sense. I do not want to equivocate them is because there are many philosophers who deny that meaning (as psychological sense) exists or is useful. Since I am not committed to identifying empirical content as mental intension or sense, those philosophers can easily read my project in their terms, using a different characterization of empirical content. However, I do not object to characterizing either meaning or empirical content this way. In fact, I am comfortable speaking of empirical content in terms of mental sense for several reasons. First, the arguments marshalled against this notion of meaning do not convince me. Quine showed that there is no analytic-synthetic distinction, and because of this concluded that philosophy has no special claim to the realm of metaphysics. According to Quine, science is the main arbiter of that realm. The question of eliminating meaning is therefore a metaphysical question, so it falls on scientists. What does modern science say about meaning (as psychological sense)?

Jerry Fodor and Zenon Pylyshyn argue that meanings are not naturalistically acceptable. They argue that mental states are representations, or sequences of representations, and they understand representation causally. They argue that this is the only naturalistically respectable theory of mind, so they conclude that there are no such things as meanings (Fodor and Pylyshyn 2015). One problem they face are concepts with empty extensions. These all refer to the same thing (nothing), and yet they are different. Examples include scientific idealizations, historical concepts like PHLOGISTON, and fictions. There is a big difference between FRICITIONLESS PLANE, IDEAL GAS, SHERLOCK HOLMES and UNICORN. The differences concern the properties we understand those things to have.

Furthermore, we do not need causal exposure to something in order to think or talk about it. We can assume with empiricists that to possess an idea requires some external event that first initiates it. But this is not the same thing as requiring causal connection for representation. We
can require causation for the existence of a thought, and we can require causation for the bare possibility of reference, but we do not have to require the reference of a thought to be the thing that caused the thought. My seeing a mathematical proof causes me to believe that a theorem is valid, but the referent of the theorem is not the symbols on the paper. My experience with horses and horns may cause me to create the concept UNICORN, but my concept does not refer to the horses and horns I have seen.

Finally, even if we eschew meaning for direct reference, we will find ourselves faced with something very much like psychological meanings whenever we discuss scientific imagination. Scientists often speak about what would, should, could, or might happen. If we insist on portraying mental states as mere representations, then we will necessarily portray scientific reasoning as making a great deal of reference to systems and states that do not exist, and wondering about what properties those systems possess. For example, suppose biologists are trying to devise an experiment to help them answer the question of the origin of life on Earth. They will manipulate representations of LIFE and ORIGIN, but they will also create and explore idealized scenarios in which life could have sparked. Then they will mentally create and explore possible experimental arrangements that could help them choose between the idealized scenarios for further testing and modelling. In the end, what matters is not having a list of all the possible but non-actual beginnings of life on Earth, but rather being able to explore a handful of likely examples of such beginnings and their properties. This explorative feature of scientific reasoning is an important step between question and solution, between theoretical structure (BEGINNING OF LIFE ON EARTH) and referent (the actual beginning of life on Earth). And since mere representation relations cannot be the link between a theoretical structure and what it represents (see Chapter 6), what is left is best described as meaning.

In any case, I leave this issue here, because empirical content can be characterized as meaning whether it is a Fregean sense or mere representation relation. According to all the above characterizations, meaning is something that is possessed by an agent, and can be increased by use of the imagination—that is what is important. So in what remains, I will treat empirical content and meaning interchangeably. And if I do this, I can marshal the resources of the philosophy of language to help explain how empirical content or meaning can be increased via the imagination.
And this brings me to the relation between empirical content and the objective elements of science such as truth, representation and knowledge.

Scientific realists argue that there is an objective reality that we approach with subsequent scientific theories. But they need not claim in addition that there is an objective meaning for each theoretical structure. It is conceivable that at the end of science all scientists will have slightly different meanings for the final set of structures, because meaning needs only be stable enough to enable communication and action. If we want to talk about something in great detail, we need more detailed meanings. But at no point do we need a determinate and objective meaning for anything. On the other hand, we hope that the meanings we manipulate and possess lead to referents for our theoretical structures in a reliable way. You and I can disagree about whether bulldogs are vicious, but we will not disagree on whether Rex is a bulldog. To explain how meaning can be inexact yet truth-conducive, we need to turn to the connection between meaning and reference.

Thought experiments solve the problem of coordination by increasing semantic content for theoretical elements. In the philosophy of language, a common view is that meaning fits between use and reference: from the use a speaker makes of their terms, we can infer their meaning. And from their meaning, we may infer their reference. If a goal is scored by my favourite team in a soccer match, and I point to the TV and jump up and down and cheer, you can determine the meaning of my words from the combination of 1) your beliefs about what just happened on the television, 2) what you think I believe happened, 3) the sounds I am making, and 4) the assumption that I am being genuine and not purposely misleading. From here, you can begin to determine the referents of the words I use, like “goal.” How exactly we make the transition from use to meaning to reference is a question in the philosophy of language that has a very long and interesting history.

And it is mirrored exactly by the problem of scientific representation in the philosophy of science, which, according to Mauricio Suárez, “is at the crossroads of attempts in analytical philosophy to come to terms with the relation between theory and the world, and in the philosophy and history of science to develop a proper understanding of the practice of modelling in the sciences” (Suárez 2010, 91). Some analytic philosophers conceive of representation as a
relation, and try to figure out the properties of such a relation, while historians and philosophers of science look at modelling practices. We might investigate scientific representation by looking at the relation between a scientific concept like \textsc{electron} and its intended referents, or we might track the history of the models of electrons (as in Arabatzis 2006). In my view, both approaches are valid: we would like to know in virtue of what something is a representation, and also how representations are used.

The problem of representation is similar for philosophers of language and philosophers of science (as long as we treat reference and representation together, as in Elgin 1996). And both questions can be seen as special cases of the problem of coordination. One very important thing to notice is that philosophers of science, unlike philosophers of language, generally ignore meaning. Perhaps it is assumed to be unproblematic, or irrelevant. For whatever reason, it is rarely mentioned in the current literature on scientific representation and models. In what follows, I will outline several influential accounts of scientific representation and consider what happens to these accounts when we bring meaning (subjective cognitive empirical content) back into the picture. To anticipate, considering empirical content between theoretical structures and experience places understanding as a mediating step between theory and knowledge. This is because while increasing empirical content or meaning is not equivalent to understanding, understanding a theoretical structure is what happens when we know what it means. I claim that this portrayal of scientific representation and epistemology is a significant advance. While meaning is usually accepted as an important mediating step between use and reference, subjective empirical content is not usually considered as an important step between model and target. Likewise, understanding is not usually considered as an important step between theory and knowledge. However, both empirical content and understanding are necessary to represent and know.

Therefore I will now reconsider accounts of scientific representation. A great deal of scientific epistemology rides on the hope of producing an account of how models represent their targets, and since the problem of scientific representation is a special case of the problem of coordination, it should also be solvable through a consideration of the mediating role of the imagination.
4 Imagination and the Problem of Scientific Representation

4.1 Similarity and Isomorphism

Reducing representation to similarity cannot be done for several well-known reasons. For example, while representation is irreflexive and asymmetrical, similarity is both reflexive and symmetrical (Goodman 1968). Isomorphism is a kind of structural similarity which suffers from the same problem (for more criticisms, see Aronson et al. 1995 and Suárez 2003). Because of these critical disanalogies, many have given up on what Suárez calls the analytical attempt to identify the representation relation between theory and world, and have turned to what are called inferential accounts of representation. Above I argued that representation cannot be identified with any abstract mapping function in a way that leaves out agential intention, so I also reject similarity and isomorphism based accounts of scientific representation. Let us turn then to inferential accounts.

4.2 Inferential Accounts of Representation

An inferential account of representation has it that representation is whatever allows us to make successful inferences through a surrogate. One example is R.I.G. Hughes’s (1997) Denotation-Demonstration-Interpretation account. Representative reasoning begins by denoting the referent of a model (usually by convention), then demonstrating something using the model, and finally interpreting the result of this demonstration back to the target system. There are several issues with this kind of account. Following Suárez (2010), I will point out two. First, it is not clear which of the various actions here counts as the act of representation. If it is the whole process, then a model does not refer until it makes some demonstration using the model, which seems strange (surely we can refer using a representation without manipulating it). If the representative act is the denotation, then this is far too simplistic. For example, it eliminates the possibility of misrepresentation. It also does not seem possible that the representative act is a demonstration or an interpretation, or any pairwise combination of the three.

Perhaps we should keep the tripartite criteria but loosen the definitions of denotation, demonstration and interpretation. Instead of denotation, we might require only that a
representative model attempt to denote something. Instead of demonstration, we might require merely that the model has the possibility of demonstrating something. However both of these modifications makes representation too cheap: using them, anything can immediately become an equally good representation of anything.

Suárez’s own choice has two necessary conditions: representational force, and inferential capacity. Representational force “is always the feature of that evolving practice that selects the target that corresponds to a particular source at any given time” (Suárez 2010, 98). Suárez claims that “the representational target is not determined merely as a result of a convention or stipulation, but must be established by the norms that govern the practice” (98). Suárez is suggesting that an evolving practice of stipulation plus the norms of science decide what a model represents. The notion that a stipulation is part of a history of other stipulations does not make those stipulations any less stipulative. And the norms of science are not specific or normative enough to pick out unique target systems. The practice of representation conceived as stipulation constrained by the norms of scientific practice is not significantly different from mere denotation. Bringing in norms of scientific practice is descriptively accurate and relevant, but the problems with representation-by-stipulation also apply to representation-by-stipulation-in-line-with-norms.

“Inferential capacity” is a property that a model has when it can be employed by “an informed and competent user to draw valid inferences regarding the target” (98). This is a more nuanced version of the demonstration step from Hughes’s account. The advantage is that it matches most cases of actual scientific representation: scientists are informed and competent users who do draw valid inferences from representations regarding their targets. However, inferential capacity also falls prey to considerations of cheapness. Any representational model will be the base of an infinite number of valid inferences about a target system in the form of derivable tautologies. If all it takes to be a representation is that it is stipulated to refer in line with the norms of science and yields valid inferences, then almost everything we intend to represent will do so successfully.

The idea behind this and other inferential accounts is that we could avoid defining representation by characterizing it as what happens when we reason through a surrogate. However as we have
seen, scientific representation is still implicitly defined whenever we attempt to define good surrogate-reasoning. And the problems with representation implicitly defined are just as poignant as when it is explicitly defined.

4.3 Exemplification

A more nuanced account of representational reasoning is Catherine Elgin’s account of exemplification. Elgin outlines her view in the following way:

Exemplification is the relation of a sample, example, or other exemplar to whatever it is a sample or example of (Goodman 1968; Elgin 1996). A fabric swatch exemplifies an available pattern of cloth; a textbook example exemplifies a logical inference. Exemplification involves a dual referential relationship. An exemplar directly refers to a property or pattern it instantiates or a relation it stands in, and thereby refers indirectly to other items that instantiate that property or pattern, or stand in that relation. An exemplar typifies an extension when it exemplifies a property common to all and only members of that extension (Elgin 2011). In highlighting, displaying, or making manifest certain of its properties, patterns, and relations, an exemplar provides epistemic access to them. By instantiating and referring to its shade, a splotch on a paint sample card exemplifies teal blue. (Elgin 2014, 224)

The dual requirements for exemplifying a property are direct representation and instantiation. Something represents directly when it is used intentionally to represent that property. And something instantiates a property when it possesses it (actually or metaphorically).

Interestingly, Elgin argues that the cognitive efficacy of thought experiments can be explained by considering their ability to exemplify properties. That is, a thought experiment both directly refers to and instantiates a property or pattern of interest. By exploring an imaginary scenario that exemplifies some properties or patterns of interest, we learn about those properties themselves, wherever else they may turn up. I will first analyze Elgin’s account of exemplification as an explanation of understanding-oriented thought experiments, and then propose a way to extend her account.
First, I will argue that we should not consider understanding-oriented thought experiments as identical with exemplifications, or functioning primarily by exemplification. Exemplification requires direct representation and instantiation. I will consider direct representation first.

The thought experiments I have analyzed are concerned with theoretical structures (which can be properties and patterns, to fit Elgin’s description of exemplification). Do thought experiments refer directly to their relevant theoretical structures? Does Einstein’s elevator, considered as a performance, refer to the principle of equivalence? Perhaps in a roundabout way we could use the elevator to refer to the principle of equivalence, but it does not do so naturally. The same can be said for Maxwell’s demon and the second law of thermodynamics. These thought experiments refer to their theoretical structures in the same way that Hamlet’s play-within-a-play refers to the regicide committed by his uncle Claudius: once you know what Hamlet is up to, you can see the play as a representation of Claudius’s act, but on its own, it is not one. The play by itself does not tell you how Hamlet feels, what he suspects, what his uncle has done, or anything besides what is said and done by the characters on stage. Equally, considering a pair of elevators does not on its own refer to the principle of equivalence.

If the thought experiments do not refer directly to their theoretical structures, is there anything else they represent directly? Again, I do not think so. A thought experiment is a performance, a set of cognitive actions. These can be made to refer to something, but on their own they are just actions. The process of imagining a sequence of mutations that lead to a vertebrate eye does not refer to the vertebrate eye or earlier eyes or selection pressures, because it is a process, which is not the kind of thing that refers.

Secondly, even if performing a thought experiment did represent the theoretical structure it was intended to explore, this would not guarantee that it also “thereby refers indirectly to other items that instantiate that property or pattern, or stand in that relation” (Elgin 2014, 224). This is because representation is not transitive. If I represent the five great extinctions of life on Earth with the digits of my left hand, I do not thereby refer to the Jackson 5 merely because the group also instantiates the property has five members. This is because representation necessitates intention. If I do not intend to refer to the Jackson 5, I do not refer to them. And when I count off the five great extinctions of life on Earth on my hand, I do not intend to refer to any other set of
things with five members. In the case of the thought experiments analyzed earlier, if
Heisenberg’s microscope was an intentional reference to the uncertainty principle, is it also a
reference to all cases of quantum uncertainty? It does not seem so.

Turning to instantiation, it is equally clear that the thought experiment itself, considered as an
action, does not instantiate the properties in question. To imagine a demon playing with particles
is not to instantiate statistical violability. Perhaps statistical violability is instantiated
metaphorically, something that Elgin allows. In this case, even if Maxwell’s demon instantiates
statistical violability, it is not really this property that we are interested in when we consider
thought experiments as tools to increase understanding. These thought experiments are not
intended to teach us something about a property or pattern, but to help us connect properties and
patterns to other properties, patterns and experiences in a way that makes them intelligible and
fruitful.

Thus the performance of a thought experiment considered as a whole does not usually instantiate
or refer to a relevant theoretical structure, although parts of the scenarios or events imagined can
instantiate and/or refer to parts of the structure. Thought experiments make use of represented
and metaphorically instantiated properties, certainly, but this is not the same thing as being one
or both of those things.

Moving on, Elgin claims that exemplars make some properties salient by instantiating them. The
red paint splotch makes its colour salient by being red. It does not make its shape salient,
although we could make the shape salient if we wanted. Of course, the paint splotch does not
make its colour salient by having it; it makes its colour salient by being a paint splotch. In other
words, we know what it is meant to do, so we know which of its properties are the salient ones.
Salience therefore depends on what we use the representation for, and we can change what is
salient as we wish. And this means that salience does not automatically result from instantiation:
it is something extra that we add. Without drawing our attention to certain properties,
instantiation does not tell us what about a representation we should notice. When salience is not
given, instantiation becomes commonplace: everything instantiates at least one property of
everything else (for example, self-identical).
A quick caveat: I do not mean to imply that Elgin is committed to the view that thought experiments are exemplifications. She claims only that they make use of exemplification, which I think is true. In the above I merely wanted to show that considering thought experiments as actions makes it clear that even if exemplification is the main epistemic tool in many or all thought experiments, this will not be sufficient on its own as an account of scientific representation or of understanding-oriented thought experiments. Exemplification itself must be explained in more detail. To see this, we can ask what would be necessary to complete the account.

To ensure a thought experiment refers and instantiates in the right way, we would need the background knowledge of the scientists to be such that it guaranteed that the properties of a thought experimental scenario referred to the properties of the scenarios governed by the theoretical structure (or the properties of the structure itself), and that it instantiated them in a way that made the relevant properties salient. But no such guarantee is forthcoming from background knowledge. Certainly background knowledge plays an important role in exemplification, but it is near impossible to say exactly what background knowledge you need, for example, in order to know what the two imaginary elevators mean for General Relativity and experience. Even if we could state it, and even if all of it was made available to an agent, can we guarantee that she would understand the equivalence principle by the thought experiment? No; we must add that the agent has and uses a cognitive ability to join elements of imaginary scenarios to patterns and properties, which recalls precisely the action that has been the focus of the last two chapters. Some bridging or coordinating action is necessary to understand the relevance of what goes on in the thought experiment in order for it to be a representation or salient exemplar of anything.

Someone might try to complete this account by having exemplification play one important role in a larger process of representation. In such a process, what is necessary (and so far left out) is a role for the agent to connect the representation to its target. More specifically, this action would identify and connect the exemplified properties that are salient in the representation to the corresponding properties in the target. To repeat, the background assumptions are supposed to determine which these are, but on their own, they do not. For example, consider a Newtonian model of the solar system that consists of point masses and equations of motion. Suppose the
point masses of this model instantiate stable elliptical orbits. We need the user to connect this property to the property *mostly stable elliptical* orbits, which we take to hold for the real solar system. This is a separate cognitive accomplishment that is not as easy and automatic as it is with paint splotches.\(^5\)

If thought experiments are mental models, and models represent their targets, perhaps this is the right account of representation to make sense of them. Let us consider more closely the action that connects the model to its target. This action takes place when scientists coordinate, for example, the level on a thermometer with a temperature, or sets of possible or actual experiences with the laws and principles of a theory. This action should not be merely stipulative. In the case of the thermometer it appears stipulative, but in reality the choice of what to coordinate with what was the result of a long and difficult experimental-contextual decision (see Chang 2004). Some actions of coordination are made easier by stipulated rules, like map legends. But map legends do not tell us where cities end, or which borders are guarded, etc. These are limitations we do not notice because we have learned what to expect from maps. Still, in every case we must make the connection between a representation and its target(s) ourselves. The fact that this is easy to do in many cases changes nothing. And again, the impossibility of complete stipulation is a good thing, as it allows models, thought experiments and other representations to surprise us (recall Hesse’s neutral analogies).

The mental acts of connection that give theoretical structures content and increase scientific understanding are not merely stipulative. The agent decides and finds out what a structure means through imaginative exploration, and through this, arrives at an idea of what it applies to. For instance, a written or verbal presentation of Einstein’s elevator thought experiment causes us to imagine experiments performed in elevators. Aspects of these imagined scenarios exemplify (refer to and instantiate) some properties of the world as described by the principle of equivalence, and these are connected to other properties thought to hold in the imaginary scenario. Most importantly, the property of empirical indistinguishability is metaphorically

\(^5\) This extension of Elgin’s account has profited from conversations with Roman Frigg that followed a colloquium he gave at the University of Toronto on September 8, 2014.
instantiated between the two elevators imagined in the thought experiment. This property is connected to the empirical equivalence of the same events that would occur if this thought experiment were physically realized, that is, if this possible world were the actual world, and to all events in frameworks that are relevantly similar to those in the thought experiment.

Now, do we emerge from the thought experiment with new knowledge concerning what it refers to? I do not think so, or if we do, this is not the interesting thing that happens. What is really remarkable is that the thought experiment shows us a kind of world. We get to explore some of the properties of this world. And when we finish, we see that our world is like that world, if the principle of equivalence is true. In other words, we emerge with an idea of the empirical significance of the principle. The performance of the thought experiment is not a reference to the principle of equivalence, and it is not a reference to the world. It is an activity that facilitates scientific understanding, which is a psychologically necessary prerequisite for the principle of equivalence to do the heavy lifting we expect it to do. If the thought experiment helps us in understanding and using the principle, it is because it forges a good connection between it and experience.

There are several things to say about this. First, I should repeat again that correlating or connecting a theoretical structure with empirical content is fallible. Heisenberg’s microscope shows that we can successfully correlate the wrong items of experience with a theoretical structure. This is a very different sort of failure than reference-failure, but is nonetheless an important kind of failure. That is, the thought experiment does not mistake photons for electrons or anything like that. In fact, all of the events that take place in the thought experiment would take place in real life if you performed the experiment; namely, you would find that the measurements you took of the position of an electron after interacting with a photon would indeed obey the uncertainty principle. Instead, the microscope fails because it assumes that the electron and photon would collide classically, which they do not.

Second, we need to expand on the aspect of the imaginative act that is non-stipulative in a thought experiment. The imaginative connection is non-stipulative insofar as it is the result of an experimental process: it emerges through trial-and-error. We make several attempts to connect various experiences, beliefs, structures, etc., with a new theoretical structure, especially when it
is first introduced. Once a connection is established as a hopeful possibility, it is adopted hypothetically and refined for communication. In classroom settings it appears that classic thought experiments stipulate which connections are to be made and how. But being given a path up a mountainside is not climbing the mountain. We have to make the connection ourselves if we want to understand the relevant theoretical structure. Luckily, thought experiments make this relatively easy. One way they do this is by building in the experimental trial-and-error process. This is why so many thought experiments consider several similar cases in a row before concluding, like the trolley problems, Newton’s cannonball, and children’s fables; they get us into an experimental mindset that primes us to pick out the correct connection when we see it.

Again, we cannot stipulate how to connect a theoretical structure to its target(s) because that connection is an action the agent must perform. Without trying out a connection and “seeing” that it works, the connection is not made. To successfully perform an understanding-oriented thought experiment means to try the proposed connection in some imaginary scenario(s) and evoke a judgment of adequacy. This judgment of adequacy is called in cognitive science literature an “aha!” moment. The theoretical structure is imaginatively filled out using existing concepts, structures, experiences, abilities, etc., and when one of these coheres nicely, subconscious brain processes trigger an emotional response in the brain that promotes the new link to a conscious level (Thagard and Stewart 2011). This emotional response is not evidence that the connection is correct, but it is reason for tentatively assuming its adequacy in order to consider it further and test it in a laboratory, computer model, discussion or with more thought experiments. Obviously, we will not always emerge from a thought experiment with an aha! moment; I confess I am still waiting for the aha! moment that was supposed to accompany Gettier cases. But understanding-oriented thought experiments that never provide a feeling of insight, whether strong or weak, should not be counted successful for any subject who is unmoved by them.

6 A Platonist might fit in here. Socrates agreed that one cannot be shown the forms, but must see them for him or herself. This is why Socrates derides rhetorical writing, not because it is ineffective, but because it is not as effective as individually tailored dialectical question and answer. In the Theaetetus Socrates calls himself a midwife to knowledge: he does not give birth to true conceptions, but rather helps others to do it for themselves.
The idea that these thought experiments will combine with mental processes to provide insight might remind us of Suárez’s notion of representative force. That notion invokes norms to explain how a representation identifies its target. My suggestion also relies on norms of science, as well as norms of individual and inter-subjective rationality. However the role of these norms is not to help establish reference, but meaning. I cannot change who ARISTOTLE refers to, but I can change what ARISTOTLE means, for me, by reading him for the first time. I cannot change what ELECTRON refers to, because this also depends on the world, but I can test in thought some different ways for me to understand ELECTRON, and indeed this is what scientists have done and continue to do (see Arabatzis 2006). I have even more control over the meaning of abstract normative concepts like JUSTICE, KNOWLEDGE, BEING, etc.

Second, I want to emphasize that the action we perform should not be considered a relation, in the sense of an abstract connection between relata. If we insist on speaking in terms of relations, then we can say that a thought experiment creates a relation between a theoretical structure and something else, but performing the thought experiment is what does the relating. This takes us back to the conventionalist solution to the problem of coordination discussed in Chapter 6. Portraying the action of connection as a relation only begs the question, because we would need another relation to tell us what to relate to what. What I am suggesting instead is a fallible, mind-dependent process of imaginative experimentation that connects as though on its own. I say “as though on its own” to allow for both subconscious and conscious mental activities. A thought experiment could connect a structure to some content “on its own” if it does so quickly and easily and/or subconsciously. However, conscious guidance is often needed. An exercise of the imagination like a thought experiment can become an action of connection by letting us vary the elements of an imaginary scenario, the relations among those elements in the scenario, or the entire scenario itself, in a way that lets us test and tentatively decide which empirical content to assign to something.

The advantage of this idea is that it bypasses the circularity of stipulative relations between models and targets. I conclude that theories of scientific representation are improved by mediating them through scientific understanding. In difficult cases, which often concern highly abstract and idealized theoretical structures which are presented in mathematical or formalized language at times of scientific revolution, determining scientific meaning can be more important
than scientific reference. This is because reference is often very easy to stipulate. We know that theoretical structures refer to the world, atoms, living organisms, etc., while precise and coherent semantic content is more difficult to specify. No one asked what quantum mechanics referred to; they asked what it meant. Further, establishing (fallibly) the semantic content of a theoretical structure is so profitable for science because it allows us to make new predictions and models, and communicate them to our peers. To repeat, what is profitable is not the thought experiment itself, conceived of as the actual imaginative scenario and process of mental connection, but the understanding that results from its use.

Before I move on to the kind of understanding produced by this ability to connect theoretical structures to semantic content via the imagination, I want to consider a few more points about mediating the focus on scientific representation with scientific understanding.

Understanding is a necessary step towards knowledge (if we ever obtain it), just like semantic content is a necessary step towards reference (if we ever obtain it). Representations are necessary in science, but making those representations and manipulating them does not directly tell us about the target system itself. Coming up with representations that exemplify the properties of our theoretical structures in a way that increases their semantic content for us is an individual matter. It is an ability that all humans have, which cannot be taught in any systematic way. But it can be improved or strengthened by regular use of the imagination. This is exactly what Kant said about the faculty of judgment, which operates to connect concepts to content.

Here is an example that illustrates Kant’s point. There are rules of chess and there are rules about how to apply those rules, but at some point if we want to play chess we will need the ability to apply rules, in general. A complete knowledge of all the rules will tell you which moves are “chessly possible” (Allison 2004, 207), but it will not tell you how to win the game. Even if we make it a rule that you must try to win, we need to learn how to apply this rule as well. And doing so is a matter of improving chess-judgment skills. We must be able to see arrangements of pieces not only as arrangements, but as opportunities to apply this or that categorization and rule, to achieve this or that end result.

What allows us to subsume particular instances under universal rules is our faculty of judgment. And good judgment cannot be taught. We cannot tell you how to recognize when this rather than
that concept in chess applies. Even if you studied all the games of all past chess masters, you would only have a long list of moves that were played in various situations, you would not know what strategies were being used by which players. That is, you could not perceive the events as being subsumed under general rules. It is up to you to master the rules and strategies and to apply them *well* in real games. That said, the faculty of judgment can be developed and guided. According to Henry Allison, “Kant insists that it needs to be practiced and that this usually requires the use of examples” (2004, 207). Examples are not necessary for bridging concepts and content. They are not the only way to improve your faculty of judgment. But they are one way.

For instance, I can tell you how to fork a queen with a knight in chess, but it will only be a very abstract description that might not make sense to you. Of course, you *might* understand immediately how forking works in a game, and in that case no example is needed. But for many people, an example will be necessary. That example might be a thought experiment. This thought experiment would not *be* the ability to apply the rule, nor would it be a relation between the rule and the instance of forking, nor a reference to the rule or to the chess pieces or their arrangements. The thought experiment is instead a *way* of building a bridge between the theoretical structure and the items that make up its semantic content. Performing it properly is what builds that bridge.

While this level of discussion is too simple to be applied to the cases considered in Chapters 5 and 6, it suffices for the point about understanding and representation. The rules and concepts of chess are like the theoretical structures of science considered above. To what do they refer? This question is often less important than the questions: what do they mean (for an agent in a context), and how can they be used? Thought experiments (or other imaginative exercises) are necessary to answer these questions. The stronger our imaginations, the better we will understand the game, and the better players we will be. This is equally true for existing structures as it is for newly invented ones. Inventing and developing a new idea is always an imaginative matter, and the imagination is no less crucial for interpreting and re-interpreting the idea as it grows, even when that idea constitutes a new scientific theory or model.

According to Galileo in his letter to the Grand Duchess Christina, those who deny the truth of his claims do so only because they are “unable to understand the arguments on the opposite side, for
these depend upon observations too precise and demonstrations too subtle, grounded on abstractions which require too strong an imagination to be comprehended by them” (1615, 199-200). The wise, on the other hand, have the power of abstraction and imagination. Galileo does not tell us, but we should assume with Kant that if we are not born with these powers, we can train them with practice. The more often we extrapolate the non-stipulated details of imaginary scenarios and bend and break them, the better judges we will be. Luckily for us, life supplies us with an infinite number of daily opportunities for this sort of reasoning (including how best to get to the dairy aisle from the produce section, or raise children, or write a paper).

Finally, I turn to the long-overdue discussion of understanding itself.

5 Understanding

Here is a general characterization of understanding that will help to situate us. Catherine Elgin writes:

We understand rules and reasons, actions and passions, objectives and obstacles, techniques and tools, forms and functions and fictions as well as facts. We also understand pictures, words, equations, and diagrams. Ordinarily these are not isolated accomplishments; they coalesce into an understanding of a subject, discipline, or field of study.

Understanding need not be couched in sentences. It might equally be located in apt terminology, insightful questions, effective nonverbal symbols. A mechanic’s understanding of carburetors or a composer’s understanding of counterpoint, is no less epistemically significant for being inarticulate.

Even a scientist’s understanding of her subject typically outstrips her words. It is realized in her framing of problems, her design and execution of experiments, her response to research failures and successes, and so on. Physics involves a constellation of commitments that organize its objects and our access to them in ways that render those objects intelligible. Understanding physics is not merely or mainly a matter of knowing physical truths. It involves a feel for the subject, a capacity to operate successfully within
the constraints the discipline dictates or to challenge those constraints effectively. And it involves an ability to profit from cognitive labors, to draw out the implications of findings, to integrate them into theory, to utilize them in practice. Understanding a particular fact or finding, concept or value, technique or law is largely a matter of knowing where it fits and how it functions in the matrix of commitments that constitute the science. (1993, 14-15)

This is a good starting point. It coheres nicely with the previous sections, and also with the idea that understanding is a unified umbrella concept that takes many different kinds of object, and both relies on and generates a wide range of skills.

To go further, I want to examine how this characterization combined with some of the above conclusions interacts with other main features of epistemology. Namely, explanation, intelligibility, usefulness, agent-relativity, and knowledge.

## 5.1 Understanding and Explanation

In Chapter 6, I characterized understanding as a three place relation between agent, object and a third thing, which is often an explanation, model or thought experiment. This relates understanding to explanation, but not necessarily, since it can be achieved in other ways as well (unless those other ways are at heart, explanations). While thought experiments and models are often explanations, they need not be. Lipton (2009) argues that explanation, while closely related to understanding, is not necessary for it. Namely, there are cases where individuals gain understanding without any need for an explanation. These may be identified by considering three epistemological goods that understanding provides which individually do not need explanations to be obtained: causal information, modal information, and unification.

Causal information is one of the main goods produced by explanations, and one of the first things we point to when we say that we understand something. We understand why the sky is blue only after we know what causes it to look blue, to us. However Lipton points out that many non-propositional means exist that offer causal information other than explanations. For instance, if we learn to ride a bike or build a computer, this depends on our coming to possess causal information that we cannot necessarily express in a propositional causal inference. And that is
what it means to have an explanation for something: to have a propositional causal argument or inference. In all cases the agent will be able to say *something* about what they understand, but that something does not always exhaust the agent’s understanding. The agent has an appreciation for the underlying causal structure of some phenomenon, but they do not have a causal explanation.

Another way to understand something is to possess modal knowledge of a phenomenon; for example, knowing that something must have been the way it is, or knowing how it could have been. Lipton begins with necessity. He gives the example of Galileo’s falling bodies thought experiment in which we come to understand why the speed of fall of bodies must be independent of weight. As a reductio argument, we know at the end that it must be that way. But if asked for an explanation of *why* a body’s speed of fall is independent of its weight, we must appeal to symmetry laws in physics. Lipton argues that establishing necessity through reductio arguments is not an explanation in the usual sense, although it does provide understanding, as we emerge knowing why something must be the case. An example from this dissertation is Einstein’s elevator, which can also be portrayed as a reductio. Since it is impossible to differentiate the two elevator scenarios in terms of possible experiences, gravitational and inertial frames can be treated equivalently.

Turning to possibility, Lipton considers examples where true facts about non-actual worlds (or false considerations about the actual world) provide modal understanding about the actual world. These cannot be explanations, since they are false. An example from this dissertation is Maxwell’s demon, which concerns a scenario that is clearly false of this world, yet provides understanding about what is possible in it.

A final way to characterize understanding is the cognitive state produced after apparently diverse phenomena have been unified. If we consider examples of unification in the history of science, we see that they are not always explanations (although many of them are). Sometimes we unify phenomena by a tacit agreement, by following what Kuhn called exemplars. The exemplars and their tacit rules concerning what scientists should do are not part of the theory proper. Yet they focus inquiry and provide structure, in a way that shows how different parts of the theory fit together. Lipton does not provide an example, but one from this dissertation is Darwin’s eye.
Readers may agree with Darwin’s general theory, but have a hard time understanding how it applies to complex cases like the eye. Darwin helps to unify many diverse phenomena under the same theoretical framework using the eye thought experiment by providing a way for scientists to investigate all traits in a similar way. The eye therefore unifies theoretical and methodological elements of the theory, even though it is not an explanation because Darwin did not intend his example to be taken as a true account of how the eye evolved.

To conclude, while understanding is clearly related to, and often produced by explanations, the latter is not necessary for the former. This is supported by the cases I have examined in Chapters 5 and 6, which serve as counterexamples of Lipton’s kind.

More generally, the cases I have presented increase empirical content, which is not always an explanatory exercise. For example, we may increase the empirical content of GOLDEN RETRIEVER by coming into contact with more and more golden retrievers. This sort of activity is not explanatory after a threshold is reached, although empirical content is still increased after that threshold.

Perhaps one reason these thought experiments seem like explanations is because semantic coordination, understanding, and explanation are all three-place relations in which one thing is related to another thing, by some cognitive agent. In a thought experiment, an agent $X$ understands a theoretical structure $Y$ by means of $Z$. $Z$ is the action of coordination between $Y$ and $W$, where $W$ is a set of experiences, memories, values, abilities, or other structures. We are tempted to characterize this as explanatory because the process increases understanding for $X$ (which is often a sign of explanation), and also because there are two options to characterize explanandum and explanans: either we explain $Y$ to $X$ in terms of $Z$, or we explain $Y$ to $X$ in terms of $W$. This tempts us to see $Z$ as an explanation, although neither $Z$ nor $Y$ is necessarily explanatory.

The success of Lipton’s arguments depends on our being able to say when understanding has been achieved. Lipton relies on some basic introspective guides to understanding, and so have I. Two main guides were introduced in Chapter 6: $X$ understands $Y$ when $Y$ is intelligible to $X$ ($X$ knows what $Y$ means) and fruitful for $X$ ($X$ can achieve a goal or can identify new possibilities using $Y$). While these have been helpful, it is time to consider them in more detail.
5.2 Intelligibility and Fruitfulness

For the purposes of this dissertation I have characterized intelligibility as “knowing what something means” (with Toulmin), and I have characterized meaning as “empirical semantic content.” Henk de Regt notes that scientific judgment is a skill, which can “help scientists to make sense of the conceptual as well as the empirical significance of explanations for the phenomena to which they refer” (De Regt et al. 2009, 8). (“Making sense” is another phrase I have used in place of “coordinating semantic content”). Machamer and Woody (1994), Chang (2009), and Eigner (2009), all discuss alternative characterizations of intelligibility and fruitfulness, which I will now use to sharpen my own.

5.2.1 Intelligibility = Usability

Henk de Regt (2009) argues “if scientists understand a theory, the theory is intelligible to them” (31). In other words, for De Regt, intelligibility is required by understanding. He defines intelligibility as “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” (31). Intelligibility is therefore not a property of something, but a value projected onto something. It is another three-place relation, connecting the agent, the virtues of a certain theory, and the explanatory models employed by that theory. In other words, intelligibility is a measure of the perceived fruitfulness of a theory to produce explanatory models, from the perspective of an agent.

This characterization of intelligibility depends on another argument of De Regt’s, namely that understanding a phenomenon is only possible when we understand the theory we are using to understand the phenomenon. To understand a phenomenon is to have an argument or explanation that fits the phenomenon into a broader theoretical framework (32). However understanding the theory, which is a necessary precondition of understanding or explaining the phenomenon, cannot be so expressed. It is the ability to use and work with the theory. He writes, “Constructing a model, and accordingly constructing an explanation, requires skill and judgment...Understanding in the sense of [having an explanation of the phenomenon] is an epistemic aim of science, but this aim can only be achieved by means of pragmatic understanding...(the ability to use the relevant theory)” (2009, 30).
De Regt is arguing that intelligibility is something to do with theories, since understanding a phenomenon can be done in the usual, Hempelian way (26). Lipton’s arguments discussed above show that explanations are not the only way to understand a phenomenon, and it is not as straightforward as Hempel claimed. Nevertheless De Regt’s characterization of intelligibility on the level of the theory deserves discussion. Here is a simple example of how it would work.

Quantum mechanics has certain theoretical virtues, and a given scientist might enjoy its abstract mathematical quality and not mind that it lacks visualizability, because a system like this coheres with her personal skill set. Another might be attracted to string theory, or quantum loop gravity, for similar types of reasons. The combination of personal skills and theoretical virtues makes the production of explanatory models of phenomena easier from some theories over others, for some people. The theories that cohere with us in this way are those we label “intelligible.” De Regt argues that we can test for this coherence by checking to see if we “can recognize qualitatively characteristic consequences of [the theory] without performing exact calculations” (33, see also De Regt and Dieks 2005). This is meant to be a sufficient but not necessary criterion for intelligibility.

I agree with De Regt that intelligibility in this sense “is not the only requirement for scientific theories, but it is an essential one if one wants to understand the phenomena” (2009, 38). The overall aim of De Regt’s analysis is to substantiate the general idea that understanding a theory means being able to use it, and that is also something with which I agree. However understanding does not seem like something that, in general, should require the valuation of theoretical virtues. Here is a simple example. If I understand Russian, I will be able to use it to create simple and useful linguistic representations without much mental effort. But I do not necessarily value the Russian language’s “cluster of virtues.” It seems possible that I could be forced to understand something (like Russian) despite feeling little coherence between its theoretical virtues and my own skill set. I think to this possibility De Regt would reply that in this case, I will only understand Russian after my skill set has grown to a level where I can use the theory to produce explanatory models. I think this reply would be correct, and I think it shows that the real lynchpin of his characterization is *ability*, not value.

One important thing to note is that De Regt has defined intelligibility in terms of what I above called fruitfulness. I think this is one way to do it, but I would prefer to keep the two separate
because I think there is more to intelligibility than the ability to use something; namely, the possession of empirical content (which may or may not be fruitful). This is discussed below.

5.2.2 Intelligibility = Unification

Peter Machamer and Andrea Woody (1994) argue that Galileo “used the balance model to make intelligible problems of dynamics as well as kinematics (as we anachronistically call them)” (214). The balance model was an Archimedean model of physics, and Machamer and Woody call it an “understanding device” (216). Galileo used it frequently, and according to Machamer and Woody, he tried to rephrase all motion as motion describable in terms of balance. Galileo begins by considering all motion as taking place in (usually the same) medium, to reduce complicated problems of motion to simple problems where the main or only variable is weight. 

Questions of motion become questions about what causes things to move (that is, become unbalanced in terms of their weight), and what could cause this balance to be restored. The famous inclined planes thought experiment is a consequence of this way of thinking about force created by uneven weight distribution and their relation to balance points (217). Seeing all motion in terms of balance unifies the treatment of statics, dynamics, rotational motion and collisions (215). Machamer and Woody argue that using a “model of intelligibility” like this one in physics classrooms will “unify [students’] thinking about motion and, at the same time, provide them with a general procedural schema for solving motion problems” (215). Therefore if I am right that intelligibility is a test of understanding, we can use the notions of these authors to show that unification and possessing a general schema for solving problems are tests for intelligibility.

Here are some of the affordances of the balance model. For one, it unifies diverse phenomena. Second, since the model is simple, Galileo feels he can understand problems better using it, and can identify clearly when an answer is reached through it. Also the balance is both a physical and an abstract mechanism; if we cannot work out the direction or magnitude of a resultant force using an imaginary balance, we can use an actual balance (although “the need to test by resorting to real physical balances was never felt to be crucial” 219).

What is even more important, according to Machamer and Woody, is the ability of the balance model to provide physically visualizable solutions to problems. These visualizations make
possible geometrical representations of problems on which Galileo later relies. The authors claim that visualizability is not necessary for a model of intelligibility, but it is “important for teaching beginning students and will aid their ability to understand” (220).

So far, what Machamer and Woody are arguing is that some abstract but cognizable models increase intelligibility by unifying distinct phenomena under a single framework, and by tying abstract questions to physical objects in a way that is clear and easy to use and perhaps also visualizable. This provides several good tests for intelligibility, although it partially conflates intelligibility with fruitfulness again.

Finally, the idea that unification is an element of intelligibility is interesting. If something unites things, it connects them, which is how I have been characterizing the understanding provided by the thought experiments discussed in this thesis. And there is reason to believe Machamer and Woody would agree with my correlating unification with bridging. They write:

> The relations among parts and the constraints on calculation are provided as much by the rules of geometry or algebra as by the physical reality of the properties of the balance, though these two are connected. In fact, it is this connection that allowed Galileo (and the rest of us before and after him) to claim that mathematics describes the parts of the world that are important for doing physics. But this is another, longer story. (221)

I hope that this thesis has told part of that story.

To conclude, Machamer and Woody argue that intelligibility is increased by unification, which I accept. They also provide a few good tests for intelligibility: if a model helps us to connect abstract problems to concrete objects or makes problems and solutions simpler, that model is a model of intelligibility (or better, an understanding device).

### 5.2.3 Intelligibility = Knowing How (to Perform an Epistemic Activity)

Hasok Chang (2009) argues that intelligibility “consists of a kind of harmony between our ontological conceptions and our epistemic activities” (64). This is in reference to the ontological presuppositions that our epistemic activities make; it is with these that our activities must be in harmony. We cannot test theories by measuring values without assuming that variables can only
take a single value. We cannot count without assuming that what we are counting is
discretizable. We cannot narrate a series of changes without assuming that something remains
continuous (against which the changes can be registered). The ontological principles just
mentioned cannot be denied if we want to perform the associated epistemic activities. If they are,
the activities become unintelligible. An epistemic activity is therefore intelligible when we do
not deny its associated ontological principle.

This leads Chang to a more general definition of intelligibility. Intelligibility is “the
performability of an epistemic activity” (75). Performability is a property of an action, and when
an epistemic action possesses it, that action is intelligible. What makes an epistemic action
performable is not-denying its ontological principle.

It seems to me that there must be more to performability than this. For example, I cannot
measure a value if I do not know what sort of thing that value is a value of. And there are a great
deal of other ethical, theoretical and technological issues that may render an action
unperformable. Perhaps Chang only wants to identify activities that are performable in principle
as intelligible activities. This means that almost all epistemic activities are intelligible, and all
those that are intelligible, are equally intelligible.

What makes this very large set of activities intelligible? I would argue that intelligible activities
are those that are intelligible for someone. There are an infinite number of activities that are
performable in the sense that their associated ontological principles are not denied (explicitly or
implicitly). But this is not what we want from intelligibility, since these activities might never be
performed or even considered. To understand something as performable, someone must entertain
the idea of performing it. I think this extra step is necessary to get us from possessing a
performable epistemic activity to understanding it. Intelligibility in Chang’s sense is therefore a
necessary but not sufficient condition for understanding.

And Chang agrees. He goes on to define understanding in terms of intelligibility. Understanding
“is knowing how to perform an epistemic activity” (75). Intelligibility is necessary for knowing
how to perform an epistemic activity because by definition, an unintelligible epistemic activity is
one that contradicts its ontological principle, making it unperformable. This makes it clear that
intelligibility is necessary and not sufficient for understanding. Chang goes on: “Understanding,
as I see it, is not some distinct quality or state of mind that exists separately from the sense of knowing how to do epistemic things” (76). Understanding is “part of all the aims of science and all other epistemic activities since understanding simply comes from performing any epistemic activity well” (79). But wait, where does “well” come from? This valuation takes us beyond Chang’s definition. Understanding is supposed to be knowing how to perform an epistemic activity (75), which says nothing about performing it well, unless “well” means “without denying the associated ontological principle.” If that is what “well” means, then understanding is identified exactly with intelligibility, which Chang explicitly denies. However Chang responds, the important point remains that intelligibility as performability is a necessary condition for understanding at most.

Chang, like De Regt and Machamer and Woody, defines intelligibility in terms of fruitfulness. Again, I disagree. It is not always the case that when something is intelligible to us we can achieve something with it or vice versa. I have argued above that one way to make something intelligible is to increase its (subjective) empirical content. I can invent a nonsense concept, and increase its empirical content, and still the concept could be unusable (perhaps because its content implies a contradiction with some relevant ontological principle). I can make something intelligible in this sense, without making it fruitful. Another example is Einstein’s early understanding of General Relativity. If we believe the story of his “most beautiful thought,” Einstein knew what the world was like, even though he could not use this knowledge. For example, he could not make predictions of the positions of objects in our solar system until the mathematics were worked out, many years later. Likewise, in the early stages of the industrial revolution, engine builders had caught on to some very interesting properties of steam. They understood how to use steam to get work done, even if it was not theoretically intelligible to them yet. They did not know which ontological principles not to deny. This is one reason to think that intelligibility and fruitfulness should not be inter-defined.

Another reason to avoid this inter-definition is that many of the actions that we usually take to be intelligible can only be called intelligible after they have been manipulated, interpreted or understood. Suppose I give a speech, and it is intelligible to you. That implies that you know what it means. Knowing what something means is the result of our having completed certain activities, but it is not those activities. You understand my speech because you hear it, make
some assumptions, and interpret it. When you say “it was an intelligible speech,” you do not mean that if you had heard it, you would have understood it. This counterfactual sense of intelligibility is a different sense of intelligibility, which is not as closely related to understanding. Perhaps something that is counterfactually intelligible is also counterfactually understandable. But this would be quite abstract. Perhaps we could say that to be understandable for some person at some time for some context is to have been understood by that person in the closest possible world where they performed the relevant epistemic actions and nothing went awry. And if we further generalize this account to an abstract or ideal cognizer, it is hard to imagine that what is understandable in this sense would be relevant to our current epistemic situation at all.

While I disagree with Chang about his specific definition of intelligibility, I completely endorse his more general project. For one, he wants to disconnect intelligibility from truthfulness or empirical adequacy. He argues that “truth is only a possible by-product of certain types of understanding” (2009, 79). This is something that I have been arguing since Chapter 5; understanding is produced by thought experiments in a way that is independent of, and sometimes necessary for, the production of knowledge. Chang calls his project “epistemic humanism,” which tries to put the human agent back into the center of the nexus of abstract and idealized manipulation of theories and models. This I also endorse, given that I am claiming active imagination is at the center of scientific understanding, which is necessary for scientific progress.

### 5.2.4 Intelligibility = Possessing Intelligible Models

Kai Eigner (2009) follows De Regt’s account of intelligibility, but in a more general way. He focuses on the intuition that understanding is “related to the **ability to use** the relevant theoretical knowledge” (273). To have understanding one must be able to use their theoretical knowledge “which implies applying it to concrete cases by means of intelligible models” (294). A model is intelligible if it satisfies two requirements. First, “if it can be used to represent the phenomenon” of interest, and second, “if its characteristic consequences can be seen intuitively” (294).

Eigner argues that the first aspect of intelligibility is really what is responsible for connecting theory and world. He argues that this is done via the creation and manipulation of representative
models. He states “it is not the case that once the model is constructed there is automatically a
connection between theory and phenomena…such a connection, without which the theory has no
empirical content and thus no empirical relevance, requires an active involvement of the model
users” (276). Eigner claims that this connection is made when we use a model to represent a
particular phenomenon. This works when we choose aspects of the model and claim that they are
similar to the target system.

Before moving on to the second aspect, there are some issues I want to raise with this
characterization. First, while I completely agree with Eigner that connecting empirical content
and theoretical structures is important when it comes to intelligibility, I disagree about what is
responsible for this connection. The bare ability to apply theoretical principles to phenomena is
not what coordinates a bit of theoretical content (model, theory, etc.) to anything. This ability is
*absolutely fundamental*, and I have provided a transcendental account of it above, where I
grounded its power in imagination. However it is *exercising* this ability that makes a theoretical
structure intelligible, not the ability itself. The structure is not meaningful simply because it has
the property of being intelligible. A connectable but unconnected theoretical structure is not
intelligible for its connectability. If we want to establish an agent-centered account of scientific
understanding, we must avoid attributing agent-centric labels to theoretical structures. When we
provide empirical content to a structure via imaginative connection, the structure gains
intelligence only in a metaphorical way. We should be more precise: it is we who gain
understanding by making a phenomenon intelligible through using the model.

Also, I would recommend caution with language like “connecting models to the world.” Eigner
is correct that one way to gain understanding is by increasing empirical content. And we can do
this by connecting theoretical structures to phenomena. My point here is already present in the
word “phenomena”: we should not be so confident that by the mere act of representation we have
gained contact with the world itself (in the sense of noumena). I am not saying that we cannot
gain such contact (I would like to remain agnostic about that). But from the discussion in this and
the last chapter, I think it is clear that we cannot simply assume that invoking a model grants us
access to the world via representation. Considering the above quotation again, notice that Eigner
is relying on an account of representation that functions much like the inferential accounts
discussed above. Although Eigner invokes similarity, which has been discredited as an
explanation of representation, and while properties are imputed to the target, there is no notion of an active imaginative connection that might substantiate the imputation. But I do not think it would be difficult to add this to Eigner’s account.

Moving on, the second aspect of intelligible models for Eigner is that their consequences can be seen intuitively. I have similar issues with this characterization of intelligibility: A theoretical structure is not intelligible if its consequences merely can be seen intuitively. They must actually be seen and appreciated. Eigner defines this “seeing” when he considers the case (from De Regt and Dieks 2005) of Ludwig Boltzman and the kinetic theory of gases. Before going into quantitative matters, Boltzman explains how thinking of a gas as a cloud of moving molecules makes sense of the relation between pressure and volume, since pressure is what is registered by the frequency of molecule impact on the walls of the container, and this frequency is increased if the container is shrunk down. Therefore it makes sense that a decrease in volume should be accompanied by an increase in pressure, for gases, in general. This is precisely the sort of example I have been focusing on in this dissertation, and I am happy to see it used by De Regt and Dieks and Eigner. The explanation for how it increases empirical content is given in terms again of the combination of theoretical virtues of the theory plus the individual skills of the scientist (as above, in De Regt). And how does this combination increase empirical content? Eigner does not say. I would answer that we can see the consequences of the theoretical structure by performing a thought experiment (or something similar), which increases the empirical content of the structure, by connecting the structure to other already coordinated structures, experiences or abilities. This connection is completed by an imaginary-experimental method that tries out connections until a satisfactory one is promoted by subconscious or conscious interaction between mental or neural faculties (memory, emotion, etc.). In this case we would not have to try very many combinations, since it is very easy to picture a shrinking box with bouncing particles whose speed remains constant.

Before I conclude this section, I should mention a few other scholars who also stress the importance of intelligibility and fruitfulness for understanding. Mary Hesse for example claims that “thought experiments have the function of rendering intelligible the language of a model” (1961, 264). Wolfgang Yourgrau maintains that “imaginary experiment has the function of unfolding the possible (impossible) extreme implications of a scientific hypothesis, so that the
hypothesis may render the theory physically convincing” (1964, 360). These are both related claims but they do not attempt to explain how the increased intelligibility or fruitfulness of a theoretical structure by means of a thought experiment is achieved, or how it is related to understanding.

To conclude then, I think that despite the many attempts to explicate understanding in terms of intelligibility and fruitfulness, these features remain best conceived of as related but separate tests of understanding. When educators test students, they try to design tests that give the highest marks to those who best understand the material. Easy questions ask students to repeat information as it was presented to them. More difficult questions ask students to apply that information to new cases. The most difficult questions ask students to do something more taxing with that information (compare, contrast, argue, relate, etc.). The students who exhibit the strongest ability to put that information to use are those that are said to understand it best. We take exemplary answers to be demonstrations of understanding. A similar structure is what makes written essays good tools of evaluation: the introductory section of the essay allows students to summarize information. Then the students have to demonstrate their understanding by doing something with that information—defending, attacking, comparing, etc. Both tests and essays are evaluation methods that check for understanding by checking first whether a student can say what important concepts mean in their own words (the student has made the material intelligible), and then second whether they can use those concepts to achieve some goal (the student has made the material fruitful). Intelligibility and fruitfulness are therefore evidence of understanding, but not understanding itself. In light of this, perhaps it would be best to think of understanding simply as an epistemically advantageous cognitive state that an agent can have relative to some phenomenon, rather than identifying it with its beneficial side-effects.

5.3 Agent-Relativity

Most of the authors discussed in this section agree that understanding is relative to an agent. De Regt, Lipton, Chang, and Eigner all argue that understanding is a cognitive achievement or ability of an agent. De Regt, Leonelli and Eigner claim that “this emphasis on the cognizant individual involves a reevaluation of the epistemic role of human agency in producing, disseminating, and using scientific knowledge. To understand scientific understanding,
philosophers must find ways to study and analyze scientific agency” (De Regt et al. 2009, 14-15). In this spirit, I would like to try to say a little more about the relation between understanding and the agent.

The main issue concerns whether understanding is an objective or subjective phenomenon. Hempel famously argued that understanding was nothing but a subjective feeling of no epistemic import (see also Trout 2002 and Yliskoski 2009). Others agree, while adding a positive spin: such a feeling can be evidence of something good to come (Gopnik 1998, Grimm 2009, Lipton 2009). Michael Scriven, on the other hand, claims that “understanding is not a subjectively appraised state any more than knowing is; both are objectively testable and are, in fact, tested in examinations” (1962, 176). But in what sense could we say that understanding is objective?

Michael Friedman presented an objective and non-contextual account of understanding in 1974. His understanding-by-unification account claimed that understanding is increased when we reduce the number of fundamental entities that we have to admit. Kitcher (1981) modified this idea by portraying understanding as what we get when we find a way to explain many different phenomena using the same patterns of argument. A more recent objective account based on unification is found in Schurz and Lambert (1994). Others like Salmon (1984) argue that we have objective understanding when we possess a causal explanation. Humphreys (1989), Woodward (2003), Machamer, Darden and Craver (2000) and Bechtel and Abrahamsen (2005) follow suit.

However there is now a growing number of philosophers that view understanding as subjective, but not in the sense of a mere feeling. This is a view held by all those who emphasize the contextuality of understanding (Achinstein 1983, van Fraassen 1980, Morrison 2009, Lenhard 2009, Leonelli 2009, Boumans 2009, Dieks 2009, van Bouwel 2009). These authors make it clear that what counts as understanding for you need not count as understanding for me. Whether something produces understanding depends on personal skill, values, background assumptions, and other features of our subjective epistemic contexts.

Elgin (2010) raises an interesting objection to this line of thought, in line with Scriven (1962). We do not think that knowledge is contextual. If something is knowledge to you but not to me, one of us is mistaken. People find themselves in many different epistemic contexts, but we do
not for that reason change the criteria for knowledge. Let us ask the analogous question for understanding: why think that understanding is relative to context? Are there contexts in which you understand that the sky is blue and I understand that it is not? That seems unreasonable. Perhaps we should say then, that a Newtonian model of the solar system did not provide understanding before Einstein; it never did, and it does not now. Elgin does not commit herself to this position, she merely asks what we should say about it.

To reply, I think one main reason many of us are not contextualists about knowledge (although some certainly are: for example Lewis 1996; Blome-Tillman 2009; Greco 2003, 2008, 2009; DeRose 1995, etc.) is because knowledge requires truth. Truth cannot be contextual because *ex contradictione sequitur quodlibet*: from a contradiction, anything follows. This is not a problem for understanding because we can have understanding without true beliefs (Elgin 2004, 2009). This is because, as Elgin argues, understanding is “non-factive.” Elgin’s imagined argument against the contextuality of understanding is therefore dissolved given the disanalogy between knowledge and understanding with respect to truth.

This does not, however, answer the question of the subjectivity or objectivity of understanding. This and the last chapter have argued that one way we gain understanding in science is by increasing the empirical content of theoretical structures. Is empirical content objective or subjective? Well, it depends. It is not subjective in the sense of being a mere (non-epistemic) psychological feeling. It is also not subjective in the sense that discipline or background beliefs or personal skills or values determine it. On the other hand, anyone can increase empirical content for any theoretical structure as they wish. The more content that is coordinated to a theoretical structure, the more an agent can be said to possess empirical content for that structure, in some sense. However, some coordinations are preferable, in the objective sense. Preferable coordinations are obtained when the chosen empirical content helps us to achieve epistemic ends, and when it coheres with the rest of our knowledge. I can attach several kinds of empirical content to $F = ma$. Perhaps I take it to mean friendliness equals masochism multiplied by anger. I can try this correlation out, but I will find that it does not cohere well with the rest of my knowledge, nor does it help me make new friends. In contrast, the Newtonian model of the solar system provides more understanding for $F = ma$ as it combines it with Newton’s other laws of motion and the universal law of gravitation, which themselves are
tied to experience, abilities and even more theoretical structures. While this increases the empirical content in a coherent and fruitful way, there is a limit. The model should eventually inspire questions about action at a distance, and we will notice that it predicts certain orbits for planets that do not obtain. One interesting thing is that since it makes us better equipped to deal with Newton’s theory, it also thereby provides us with the means to say that it is false. After all, we cannot disprove what we do not even understand.

This raises an interesting issue. Does coordinating a theoretical structure to the “wrong” empirical content still lead to understanding? What makes empirical content the wrong content? In other words, at what point does idealized empirical content lead to misunderstanding? I do not think the answer has to do with the truth or accuracy of the content itself. As we saw several times above, connecting theoretical structures to abstract and idealized content is very often a good idea, but that content will not be accurate as a representation of the target system (because for example, real planes are not frictionless). On the other hand, good connections will not take us completely away from what we know about the world. This is why $F(riendliness) = m(asochism)a(nger)$ does not work.

There is an idea that I think is helpful here: Elgin (2004) argues that epistemic acceptability requires only that our sets of related beliefs together be “true enough,” rather than demanding that each be true. Supposing that the goal of science is truth, this does not mean that every part of every scientific activity must have truth as its goal. Equally, the goal of increasing the empirical content of a theoretical structure is understanding it. We can see if this goal has been achieved when the theoretical structure has become intelligible to the agent (they possess the meaning of the structure accepted by the relevant community) and when the agent can use the structure to achieve his or her epistemic goals. Whatever empirical content allows the agent to get to this state is good empirical content, from the perspective of seeking understanding.

Achieving the above-mentioned goals is objective in the sense that the outcome of experiments and exams are objective: because they are tests. You cannot merely stipulate that you understand $F = ma$, you must be in the state that would allow you to pass the tests of the relevant linguistic community. In some cases, there is no relevant linguistic community, and you have to develop and administer the test to yourself, as Einstein did. This is also what happens if there is a relevant
linguistic community, but you do not have access to it. The best we can do in those cases is consider intelligibility (check for consistency) and fruitfulness (check for new abilities).

Now, if we allow for multiple end goals, understanding becomes relative to the goal, and in this sense, understanding is subjective. I understand my government’s laws in the sense that I stay out of jail. But I do not understand the laws in the sense that I can prosecute high-profile defendants in fraud trials. Here “understanding the law” is relative to the goal. The same person (me) both understands and does not understand the law. However, once a goal is established, it becomes an objective matter whether or not I understand the law: I do in the first context if I stay out of jail, and in the second if I have the ability to prosecute high-profile defendants in fraud trials. Analogously, some connections to empirical content will count as understanding according to some epistemic goals, and not according to others. This is not an issue in science, because the overall epistemic goal is always to achieve the deepest understanding possible. Given this shared epistemic goal, increasing understanding is an objective pursuit.

An agent’s understanding of a theoretical structure is therefore evinced by the ability to pass certain tests, and arriving at the state where we could pass these tests is part of what it is to understand. This is how we constrain our imagination in performing understanding-oriented thought experiments. We allow ourselves to consider strange and outlandish cases that can depart from reality as much as we like, as long as they help us make connections to other structures, experiences, knowledge and abilities that allow us to pass our tests. Modifying Elgin’s notion of “true enough,” we can say that empirical content must be intelligible enough and fruitful enough. The intelligibility requirement demands coherence with our other beliefs, abilities, values, etc. The fruitfulness requirement demands the appearance of new abilities that help us reach our epistemic aims (for example, communication, measurement, model-building, problem-solving, etc.). This is why Heisenberg’s microscope, while connecting a theoretical structure to kinaesthetic sense and experience, is neither intelligible enough (it leads us to believe that the reason for uncertainty is technical-instrumental and not formal, and therefore it is inconsistent with our other beliefs about quantum mechanics, something Bohr noticed immediately), nor fruitful enough (using the connections it forges will cause us to miscommunicate with other physicists and build poor explanatory models of other complementary variables like energy-time, spin on different axes, etc.).
This way of conceiving understanding preserves the idea that understanding and knowledge are closely related—not in the sense that each requires truth—but in the sense that both can be evaluated objectively, at least in science. Still, this relationship must be analyzed further.

5.4 Understanding and Knowledge

Whether understanding is a kind of knowledge is a contentious topic. Lipton (2004) argues that understanding is knowledge of causes. Others assume that whatever it turns out to be, it will be knowledge (Achinstein 1984, 23; Kitcher 2002; Salmon 1989, 134-5; Woodward 2003, 179). Stephen Grimm notes that the characterization of understanding as a kind of knowledge is “rarely doubted” in philosophy of science, but that “every major epistemologist who has thought seriously about the nature of understanding” denies this (2006 516). Some of the main reasons for this are that understanding does not require truth, cannot be Gettiered, and is transparent (where knowledge is not). Grimm argues that each of these is mistaken.

Grimm makes his case using thought experiments. 1) You discover that your fridge is not working. You see that it is unplugged. Now you understand why it is not working. But actually, its fuse blew and someone had subsequently unplugged it for fear of fire. So you did not really understand why the fridge was not working. 2) Your roommate was drunk last night and wakes up to see that your vase is smashed. Your roommate blames the dog, who in fact did break it. You do not really understand that the dog broke it, because it is sheer epistemic luck on the part of your roommate that she got the story right. Using cases like these, Grimm claims to show that truth matters for ascriptions of understanding, which can be Gettiered.

There is certainly a sense of understanding for which Grimm is correct, but there is also a sense of understanding invoked in each example that is overlooked. Taking up the sense of understanding that I am developing in this thesis according to which understanding involves coordination, you understand perfectly well that the fridge was turned off, and that the vase was broken, and you understand also the ideas of fridge-unplugging and dogs-bumping-into-things. What you do not understand, if Grimm is correct, is why what happened, happened. For cases like these, Grimm seems to ally with Lipton, who argues that to understand is to have causal knowledge. Here is another way to put it. In these cases, Grimm is concerned with understanding-why questions, and in this thesis I am concerned with understanding-what
questions. Understanding-what is absolutely essential to science. It concerns the meanings of theoretical structures, which we fix and develop, and it makes possible our basic ability to work with concepts. Now, is understanding-what a sort of understanding-why?

It is easy enough to establish that understanding-what is not a sort of understanding-why. Grimm is correct that the latter asks for causal explanations of states of affairs, can be Gettiered, depends on epistemic context, and requires truth. As we saw above, understanding-what is not necessarily related to explanation. Second, understanding-what cannot be Gettiered. This is because understanding in this sense requires only the connection of a theoretical structure to something that provides empirical content, making the structure intelligible and fruitful. In this sense, scientists who understood electrons as an elementary quantity of electricity (Arabatzis 2006, 70) achieved understanding, despite meaning something other than what we mean by ELECTRON. Their theoretical structure played a certain role in their web of concepts, norms and practices, and was meaningful for them. Electrons are now understood as something else, which is also meaningful and fruitful. Therefore understanding-what does not require causal explanations, cannot be Gettiered, and does not require truth.

Having distinguished between understanding-why and understanding-what, I must also ask if the latter is a type of knowledge. Here is my proposal. Let us assume that knowledge is something like true, justified belief. In other words, the members of the extension of KNOWLEDGE are all beliefs that enjoy certain properties. They are true, believed by, and justified for, some agents. Another way to put it is that knowing is a certain kind of privileged relationship that an agent can stand in with respect to a belief. These are equally legitimate and we should consider them both. Can understanding as I have characterized it in this dissertation be portrayed as referring to a set of beliefs with certain properties? No, because what we understand are theoretical structures, which include abstract and idealized models. Beliefs must be had by someone, but models cannot be possessed by anyone (someone who considers the billiard ball model of a gas “has” that model no more than they did before they thought about it). We can understand a model, but we cannot know it (except in the sense of acquaintance). What about the second characterization: is understanding a privileged relationship between an agent and a belief? Again the answer is No, because understanding is not limited to beliefs.
I want to be clear: some kinds of understanding may imply kinds of knowing, especially knowledge-*how* and knowledge-*what* (knowledge by acquaintance). Understanding in the sense outlined in this thesis is not the same as or translatable into any of the standard characterizations of propositional knowledge (knowledge-*that*). This is not to deny that understanding and knowledge are related. For example, understanding-*what* usually manifests itself through intelligibility (which might be best expressed as knowing-*what* something means) and fruitfulness (which might be best expressed as knowing-*how* to do something). Since I do not identify understanding-*what* with either of these signs of understanding, this merely shows that that understanding is related to several different kinds of knowledge because having understanding implies having (some types of) knowledge.

This is an important point so I will emphasize it again. In Chapter 5, I claimed that certain revolutionary thought experiments could fulfil two different epistemic roles. Conceived of as actions, they could connect theoretical structures to other structures or existing knowledge or abilities, etc., and this would increase empirical content and thereby produces understanding. Conceived of in terms of their all-things-considered propositional upshots, which were made possible by the understanding they provided, they could be characterized as evidence for or against theoretical claims and thereby stand to increase our knowledge. EPR, for example, is an exploration of the connection between the quantum mechanical formalism and what we should expect to find given quantum entanglement. The all-things-considered upshot is that the supporters of the Copenhagen interpretation had to either give up locality or realism. But this is an *interpretation* and *application* of what was found in the thought experiment; namely, if we want to connect quantum mechanics to experience in the way Bohr suggests, we will have to do it in an awkward way, or posit hidden variables.

This explains how some thought experiments can be both good and bad. A thought experiment can be good for exploring a theoretical structure and doing so in a way that increases understanding. However, the all-things-considered upshot of the thought experiment could be incorrect. For example, Schrödinger’s cat correctly shows that the Schrödinger equation makes possible macroscopic instances of superposition. However, Schrödinger took this thought experimental conclusion to be a counterexample to the Copenhagen interpretation, which has since found a way to embrace it. Something similar goes for EPR; its authors connected quantum
mechanics to certain possible experiences concerning entangled particles, and concluded that this connection was theoretically unattractive. However, it turned out with the Bell inequalities and the work of experimenters like Alain Aspect that rejecting local realism is actually better (for now) than adopting a hidden variable theory.

What I am suggesting is that Schrödinger increased his understanding when he tied the theoretical structures of entanglement and superposition to a cat in a box. But he did not achieve knowledge when he used this understanding to attack the Copenhagen interpretation. Einstein, Podolsky and Rosen achieved understanding through their thought experiment, but they also did not get new knowledge. Now compare this with Heisenberg’s situation. His thought experiment increases the empirical content of the uncertainty principle. But it did not do so in the right way, and therefore understanding was not achieved. This thought experiment was not used as an argument for or against the Copenhagen interpretation, but we can safely assume that if it were, it would also have failed for invoking at least one false premise (for example, the classical interaction between photon and electron). On the other hand, Maxwell’s demon provides both understanding and knowledge, it seems. The same can be said of Darwin’s eye. Evolutionary theory really can handle threshold organs like the eye.

There are therefore several ways a thought experiment can go wrong, and this is at least partially due to their having the different epistemic goals of knowledge and understanding. Using a thought experiment to achieve understanding goes wrong when intelligibility or fruitfulness does not result. Using a thought experiment to increase knowledge goes wrong when the understanding it relies on is faulty, if the inferences it makes are invalid or not cogent, if there are false background assumptions combined with the thought experimental conclusion to infer to the all-things-considered-upshot, etc. This is not to say that all thought experiments that increase knowledge also increase understanding: there are many very simple thought experiments (usually counterexamples) that prove a contention wrong and thereby increase knowledge without increase understanding. These will still rely on previous understanding, but they need not necessarily increase understanding. But the cases considered in this dissertation, which I think form a substantial and interesting class of thought experiments, are such that they produce knowledge only if they also (first) produce understanding. Understanding and knowledge are
therefore very closely related and they have overlapping epistemic aims, but not all kinds of understanding are equal to all kinds of knowledge.

6 Conclusion

The foregoing discussion yields several important conclusions. First, I showed that neither by the invocation of models nor neuroscience can we straightforwardly explain the power of the imagination to create scientific understanding through thought experiments. I provided instead a transcendental justification of the role of the imagination in creating scientific understanding.

Then I discussed the nature of the understanding produced. This understanding is much closer to scientific meaning than scientific reference, which is an idea that has almost no voice in the philosophy of science. To see how it could be developed, I placed it in the context of scientific progress by considering it as an intermediary step between scientific representation and target systems. To do this I examined several influential accounts of scientific representation, and showed how the imagination could mediate the gap between our models and our experience of their target systems.

Finally, I turned to the literature on scientific understanding. First, my thesis supports Lipton’s argument that understanding does not require explanation. Second, it suggests that intelligibility and fruitfulness are not constitutive of understanding, although they are signs of it. A third consequence is that understanding can be both subjective and objective (although not assuming the same epistemic aims), depending on whether its aim is fixed. In science it is usually objective. Fourth, I showed that there are different sorts of understanding just like there are different sorts of knowledge. Understanding cannot be equated with propositional knowledge due at least to the role played by models in understanding. Yet, for the thought experiments discussed in this dissertation, understanding appears to be necessary for knowledge. Finally, I traced several important ways thought experiments can fail, separated in terms of failure to achieve their different epistemic aims.
In Chapter 1, I introduced what has been called the paradox of thought experiments. It revolves around the “puzzling fact that thought experiments often have novel empirical import even though they are conducted entirely inside one’s head” (Horowitz and Massey 1991). I showed that this paradox may be expanded into several different puzzles because “novelty,” “empirical import” and “entirely inside one’s head” may be interpreted in many ways. John Norton, for example, asks how thought experiments justify empirical knowledge (import) that was not previously justified (novel), without recourse to any new sensory experiences that could justify that knowledge (inside the head). His answer follows the form of his question: the mind justifies the new knowledge. And the way the mind does this is by inference. In general, this is probably right. But Norton identifies thought experiments with a specific notion of arguments, and not all inferences are arguments. It may be true that all inferences have some form or other, but as Norton himself argues, at least inductive inferences do not succeed because of their form (Chapter 3). Second, supposing that thought experiments are always exactly as justified as their reconstructed arguments, this is not a surprising fact that is best explained by their identification, because Norton defines an argument as whatever exhibits a form that is categorized as good by any present or future logic (Chapter 2). Since logic is in the business of capturing inferences that we think are successful, every successful cognitive mechanism automatically becomes an argument. This is circular. Finally, Norton claims that insofar as thought experiments are justified, it is because they are logical. However I argued in Chapter 3 that logic does not justify. It merely tells us more about what we already take to be justified.

Suppose we agree with Norton’s motivation, however, and want to avoid rational intuition in epistemology. If thought experiments are generating new empirical evidence by the mere
exercise of the mind, and Norton’s account does not show how this happens, then we should turn to cognitive science. Paul Thagard is a philosopher and cognitive scientist who argues that thought experiments should be avoided. They are dissociated from experience, making them a priori. And they aim to establish necessary truth, which cognitive science shows to be a notion that is “just as empty as the notion of the a priori” (2010a, 38). According to Thagard, both the method and aim of thought experiments should be buried in the Cemetery of Philosophical Concepts. While I reject Thagard’s arguments against the use of thought experiments, I accept both the results of his studies on the mind, and his challenge of trying to discover just what role thought experiments do play in science. This brings us to the main contribution of the dissertation.

Characterizing the role of the imagination in thought experiments as producing understanding sheds light on the historical cases considered in Chapter 5 and the empirical studies considered in Chapter 6. I argue that: 1) Thought experiments can perform several epistemic functions, and they can perform more than one function even for the same person. 2) In those cases where understanding is increased by a thought experiment, it is often because that thought experiment bridges some theoretical structure(s) to existing knowledge, other theoretical structures, experiences, abilities, emotions or values. 3) This bridging solves a general version of what van Fraassen calls the problem of coordination. I argued that solving this problem is a real scientific desideratum, and that scientists have historically attempted to solve it using thought experiments. 4) I substantiated this idea by considering the imagination as a key component in building these bridges in a way that cannot be questioned epistemologically due to its being presupposed by that very questioning. 5) I considered the relation between understanding and knowledge and concluded that understanding theoretical structures is necessary for the possibility of scientific progress. What makes that understanding possible is not the thought experiment itself, but an active exercise of the imagination. The thought experimental format, however, aids the imagination.

One thing to emphasize again is that my sketch of how thought experiments increase understanding does not contradict any of the existing accounts that aim to justify empirical knowledge produced by thought experiments because I am not here concerned with how thought experiments produce knowledge. Thought experiments can be directly evidential (as in Brown,
Hopp, Norton, Elgin) or they can provide evidence for a premise in an argument (De Mey, Hägqvist, Williamson), or they can be a mental model that “mobilizes” unarticulated knowledge (Miščević, Nersessian, Gendler). None of this contradicts what I have claimed about this other interesting thing that thought experiments can do, as no one in the literature denies either that thought experiments can produce understanding, or that understanding is an important step in the production of scientific knowledge. Here is a quotation from Norton that demonstrates the point from both extremes of the debate about thought experiments:

Brown…argues in ‘Why Empiricism Won’t Work’ that a goal of a thought experiment can be to produce understanding rather than novel conclusions. He points out that this goal is not incorporated in my account, and he is right. It is not. Yet I do not see that he has established that there is any incompatibility between my account of thought experiments as arguments and the possibility that thought experiments may aid understanding. Indeed, it is the practitioners’ lore in the mathematical sciences that one does not really understand this or that corner of some science until one has put sweaty hours into deriving results within it, even if one already knows the results. This is a clear instance of understanding being enhanced by the execution of arguments (derivations) to already known conclusions. To pursue the matter any further would require a more precise notion of ‘understanding’ than Brown or I have offered. (1996, 339, note 11)

Since Norton and Brown both accept that the production of understanding is a legitimate goal for thought experiments and that it is worthwhile to discuss it, my work in the last half of this dissertation could be added to either type of account without loss.

1 Implications

There are a few very general implications of this thesis that I would like to explore. They concern theory interpretation and proliferation, thought experiments and computer simulations, and the re-introduction of semantic interpretation into the epistemology of science.
1.1 Theory Interpretation and Proliferation

Many of the most cited examples of scientific thought experiments are drawn from early debates in quantum mechanics. One reason is that this is a period with an especially large and difficult gap between the new conceptual relations and what these relations might mean for experience and theory. I do not think it is a coincidence that thought experiments were used quite frequently during this period: giving meaning to our theoretical structures is always cognitively necessary in science. We may progress quite far without making sure our variables have well-defined meanings, but at some point we must stop before we go on, or run the risk of turning into pure mathematicians. Science cannot be done purely a priori; we need empirical content, and sometimes this means waiting for understanding to catch up to knowledge. At other times, as in the industrial revolution, understanding precedes knowledge. (There, the understanding was mechanical-technological. But in other cases it is physical-theoretical, as with Einstein and General Relativity). There are many ways to generate this understanding, and one way, I argue, is by thought experiment. Scientists know this, and they sometimes use thought experiments explicitly for this purpose. Mara Beller writes, “most physicists, Bohr and Heisenberg included, wanted more: some feeling of understanding, of illuminating, or explaining the kind of world that quantum formalism describes. The need for this kind of metaphysical grasp is not merely psychological but social as well—the power of a successful explanation and the power of the effective legitimation and dissemination of a theory are connected” (Beller 1999, 107).

If Beller is right, thought experiments can make possible one of the strongest motivations to believe a theory: the social, psychological legitimation we experience when a theory gives a feeling of understanding and illumination. Scientists sometimes realize this and take advantage of it. This is why we see thought experiments as such a powerful tool in theory choice and proliferation. This is not always because they incite or justify the revolution, as Kuhn argued, but because they provide higher-level understanding that humans crave. Providing an imaginative interpretation of a theory can be a way to get others to accept that interpretation, and therewith the theory.

Likewise, a common reaction for those who oppose a new and competing theory is to try to find a counterexample. These are cases where the theory does not apply, or that the theory cannot
explain. Searching for counterexamples is itself an attempt to explore the connection between the new theory and experience, and show that the proposed connection cannot be made, or if it can, the physical or theoretical correlate is implausible. This makes sense given my characterization of thought experiments, whose role in scientific public marketing now becomes clear. If a theory has been developed in great theoretical or mathematical detail, but has not yet caught the eye of the greater scientific community, perhaps it is time for some thought experiments which would assist in funding and public image. We should not forget, however, that thought experiments also work well in smear campaigns against competitors. Late night infomercials on television encourage you to imagine yourself in some uncomfortable situation, from which only the Brand New Shining Product can save you. Thought experiments are very powerful tools of advertisement. And recognizing this power highlights a danger that was hinted at in the last chapter. High-level or general understanding is one of the goals of science. Since thought experiments can provide this, they might be used (intentionally or not) to deliver such understanding falsely. Heisenberg’s microscope is an example. While it does provide a way to understand and visualize the uncertainty principle, it has been harshly criticized for doing so in a misleading way (for example, Roychoudhuri 1978).

This is an interesting issue, because general understanding, while a desideratum, might not always be achievable. Our cognitive limitations do not always match the limitations we find in theoretical structures. Perhaps it has already happened in science that we have abandoned a good theory for a rival that was more easily intuited and understood, although false. Physicist Paul Dirac “regards models, images, pictures not only as redundant, but as dangerous. As long as the formalism and experimental results dovetail, theoretical physics has achieved its task” (quoted in Yourgrau 1967, 866). But we cannot deny that intelligibility does have strong effects on science. The Aristotelian theory of motion including natural places for the five elements strongly appeals to the imagination: we easily imagine how all the parts hang together, and we know how to apply the theory to many different phenomena. This might be one of the reasons it was dominant for so long. And this possibility is one of the strongest reasons someone might be skeptical about the usefulness of thought experiments: some false things are easily imagined and therefore dangerously powerful, while some true things cannot be imagined, and we may go astray either by rejecting them for being unimaginable, or by trying to make them imaginable, even with the best intentions. On the other hand, while it is true that passing into the atomic domain or higher
dimensions blinds our ability to visualize properly or at all, this does not stop us from focusing on aspects of those systems that we can imagine. The entities that make up our world display a multitude of interesting properties, many of which stand in relations that can be visualized apart from the rest. Of course, the more complicated our theories become, the more careful we must be with our imaginary examples. Pursuing this question further will help us to map out the difficult grey area between good and bad uses of understanding-oriented scientific imagination.

1.2 Thought Experiments and Computer Simulations

The work of the last two chapters bears on the recent claim of Chandrasekharan, Nersesian and Subramanian (2013). These authors claim that thought experiments will soon be replaced by computer simulations. Perhaps these authors are correct that some of the roles that have traditionally been played by thought experiments will increasingly be played by computers. However, this cannot be the case for the role that I have been emphasizing in this thesis. This is because computer simulations cannot be used to link theoretical structures to empirical content on our behalf. For one thing, there is no database that a computer could search through in order to find an appropriate example that makes sense of a theoretical structure in a way that increases its empirical content, at least not in the foreseeable future. As far as I know this is not something that computer programmers have even begun to consider as a possibility. Emulating regular human rationality in a computer is difficult enough without asking it to think metaphorically. Second, the connection that thought experiments make must be made by the agent in question. Even if we suppose that a computer could perform an understanding-oriented thought experiment, this would not remove the necessity for humans to perform those same (or functionally equivalent) thought experiments themselves—unless science has been handed over to computers completely.

1.3 Semantic Interpretation and the Epistemology of Science

In Chapter 7 I suggested that some of the same tools we use to understand how humans interpret each other can also be applied to understand how humans interpret the world around them. Many times when we enquire, we use an experimental method, and this goes equally for linguistic interpretation (Stuart in press). Focusing on the linguistic practices of scientists will not provide
a complete epistemological account of science. But no such account will be complete without looking at the foundation of interpretive practices in general as an experimental method that involves the use of the imagination. I have tried to show in this dissertation that such a focus is useful in analyzing thought experiments, but there is much more work to be done. Elsewhere I have argued that this method can be applied to discuss conceptual analysis (Stuart in press). In future work I would like to ask how this methodology, which might be called experimental interpretationism, can illuminate the remaining steps in scientific representation.

Even within the literature on thought experiments there is much more that can be done from this angle. Since thought experiments employ a narrative structure, we could apply the resources from understanding fiction to understanding thought experiments (Ichikawa and Jarvis 2009). There is also the role that empathy plays. In linguistic interpretation empathy matters a great deal for achieving understanding. Eigner (2009) and Koster (2009) argue likewise that in the social sciences of psychology and history respectively, empathy matters for gaining theoretical understanding. I would hazard that increasing empathy will increase our ability to imagine and thereby understand theoretical structures in science.

2 Relation to Other Views

The second half of my thesis is not in direct conflict with any account that attempts to explain how thought experiments produce new empirical evidence. All I have shown is that understanding-what questions are methodologically prior to understanding-why questions (which may be knowledge questions, as in Grimm 2006), and that some sort of semantic step must mediate the connection between theory and reference, which has been mostly ignored in the literature. Brown could be correct that the mind grasps truth directly via some kind of rational intuition. I merely add that we must understand the window-latch mechanism before we can open it and gaze into heaven, and that thought experiments can help with this. Someone who supports the argument view can ignore what I have said about the imagination and understanding, and focus entirely on the application of thought experimental conclusions to theory in an effort to produce knowledge. The etiology of the beliefs that arise from a thought experiment can be black-boxed.
Therefore in this section I will focus on the two accounts that I think are closest to my own, as these are the ones in the best position to provide evidence against me. The first is the Kantian account of Marco Buzzoni. The second is Catherine Elgin’s.

2.1 Marco Buzzoni

Marco Buzzoni has been developing a Kantian account of thought experiments for some time now (see, for example, 2008, 2010, 2013a/b), and it is in many ways the inspiration for the account presented in Chapters 6 and 7 of this dissertation.

I will first say why I agree with Buzzoni’s account over Fehige’s rival Kantian account of thought experiments (Fehige 2012, 2013). The main point of contention, according to Buzzoni, is how to view the nature of the Kantian a priori with respect to thought experiments. Fehige argues that we should understand thought experiments as revealing the contingent and relative constitutive principles that underlie scientific practice (as in Friedman 2001, 2002). Buzzoni argues that we should see thought experiments as themselves being a priori in the sense of being universal, necessary, and devoid of content. Fehige follows a material reading of the a priori according to which a priori principles give content to science, and are not universal or necessarily true. Buzzoni follows a functional reading of the a priori according to which the capacity of the mind to reason counterfactually is necessary for all scientific action, and has no specific content of its own. For Fehige/Friedman, a Newtonian would take Euclidean geometry and Newton’s laws of motion as a priori. These have specific content. After Einstein, these were replaced with a new set of mathematical and physical principles with different content. For Buzzoni, on the other hand, the same thing is a priori for Newtonian and Einsteinian physicists: the capacity of the mind to reason counterfactually.

What I have argued in this dissertation is in line with Buzzoni’s reading of the a priori, although I think it can recover some of the intuitions of Fehige’s account as well. What is crucial to the production of scientific understanding is the coordination of theoretical structures with empirical content. Fehige and Friedman would argue that many of the theoretical structures discussed in this dissertation are a priori in the sense of being constitutive of a certain scientific paradigm. However those structures are not useful unless they are coordinated with some empirical content. This is precisely what Friedman argues:
This fundamental problem is that the mathematical representations employed in modern physics have become increasingly abstract in relation to concrete sensory experience…For precisely this reason, however, there is a new problem of somehow coordinating our new mathematical representations with concrete sensible experience before we are even in a position to be fully explicit about what our new physical theory actually says. (2001, 76)

Friedman argues that Newton, for example, solved this problem by formulating the laws of motion in the *Principia* (2001, 76). According to Friedman, “the laws of motion…therefore function as what Reichenbach…aptly calls coordinating principles (axioms of coordination). They serve as general rules for setting up a coordination or correspondence between the abstract mathematical representations lying at the basis of Newtonian physics…and concrete empirical phenomena to which these representations are intended to apply” (76-77). I approve of Friedman’s project—we must ask how semantic content gets provided for the theoretical structures of science. However I have been arguing in this thesis that rules or principles of coordination cannot be the answer. Friedman has chosen the conventionalist solution to the problem of coordination, which falls prey to Cassirer’s objection (Chapter 6): we cannot coordinate theory and experience by principles or rules, because those principles and rules must themselves be coordinated. What is necessary for coordination is not more structures, but an action performed by the agent. And what makes such an action possible is the imagination. The acts of imagination that are a priori in this functional sense are necessary, and have no specific content.

If Fehige admits that such active psychological connections are necessary for increasing the semantic content of scientific structures (theories, models, laws, concepts, etc.), he might also allow that an agent-centric epistemology will ground the power we have to make these connections in a cognitive ability of ours. Fehige might be correct that thought experiments can help us to reveal the relative and contingent constitutive principles in science. He can also be correct that without these principles science cannot proceed. However what is more fundamental to the creation of those principles, how we understand them, and what enables us to use them, is the functional a priori, which does not and cannot change.
If I had to offer a criticism of Friedman’s Kantian philosophy of science, it would be that Friedman sees science as a system of reductionary relationships of semantic content, where all semantic content comes from a small basic set of principles. I do not think this is necessary. Scientific principles do not have to be given semantic content only through more fundamental principles. Yet they must at some point be given content through cognitive action. This is why a material a priori considered in terms of principles expressible in language cannot be as fundamental as the functional a priori, which enables this action of connection.

Now I will present the specifics of Buzzoni’s account and some criticisms and replies, in order to show in particular how his account is related to my own. Buzzoni usually begins by drawing attention to the relationship between laboratory experiments and those carried out in the mind. He charges existing accounts of falling into the Scylla of blurring thought experiments into laboratory experiments, or the Charybdis of removing their connection completely (Buzzoni 2008, 80). For instance, Brown’s Platonism makes thought experiments functionally identical to laboratory experiments because both provide direct access to the truth. In both contexts, we work our way through some set-up and see nature’s answer to our inquiry. Norton’s empiricism on the other hand completely separates thought and laboratory experiments; one provides empirical information, and the other rearranges it. Buzzoni explains that there are in fact two levels of analysis from which we should consider the relationship between thought and laboratory experiments, in order to see how they are both the same and different. The first is the concrete level, which is the level of analysis that considers the content of the experiment and the reason it is performed. On this level, thought experiments are equal to laboratory experiments because both concern the same features of the world, ask a question to nature, vary a variable, and anticipate an answer by moving inductively or deductively to a conclusion about the target system. But from the transcendental level of analysis, which asks for the conditions of the possibility of our scientific activities, thought and laboratory experiments are different because the former, by channeling the mind’s power to think counterfactually, are the precondition for the latter. To be clear: on the transcendental level we find an important asymmetry. Thought experiments are the logical prerequisites of laboratory experiments because thought experiments give meaning to the questions that motivate laboratory experiments. Without our ability to distance ourselves conceptually from the actual situations in which we find ourselves we would not be able to think of how things might be, and in this case we could not generate questions or
hypotheses about what is. Not until we can entertain different hypotheses about what the world is really like (as opposed to how it merely appears) does experimental practice become possible.

To consider the relation between Buzzoni’s account and my own I will first consider what he says about the concrete level of analysis, and then the transcendental.

I agree with Buzzoni that thought and laboratory experiments can perform the same functions on the concrete level. What I have done is add one very important function to the list of functions that they can both perform. Namely, both thought and laboratory experiments can be used to help agents make sense of theoretical structures by connecting them to experience, other structures, abilities, etc. In addition, both kinds of experiment have the same dual nature; that is, their results are useful not only for providing understanding, but they may also be brought to bear on a theory or claim to provide knowledge. And again I agree with Buzzoni that on the transcendental level, thought experiments and laboratory experiments are different. What makes the laboratory experiment possible for Buzzoni is the kind of counterfactual reasoning we see in thought experiments. I prefer to focus on the imagination rather than the ability to reason through counterfactuals because I do not think that we must necessarily “counter the facts” in order to produce the understanding necessary for scientific progress. But this might only be a terminological difference.

Now, I want to turn to Buzzoni’s account of the relation between thought and laboratory experiments on the concrete level of analysis. For Buzzoni, thought experiments and laboratory experiments are “indistinguishable from the point of view of their logical-formal or empirical concrete (technical-operational) intentions” (2008, 117, emphasis removed), or in other words, “all thought experiments must be thought of as translatable into real ones, and all real experiments as realisations of thought ones” (110, emphasis removed). Is this too strong? Buzzoni argues that this relationship is a consequence of combining the principle of empiricism and Buzzoni’s technical operationalism. The principle of empiricism merely states that any “doubts or misunderstandings in physics can in principle always be identified and sooner or later eliminated by means of experience” (2013, 279). This he takes directly from Kant (1781, B 452-453, AA III 292, lines 27-31).

Buzzoni’s technical operationalism
takes as its methodical starting point concrete human agents who are—through their bodies—always already in operational or technical interaction with the surrounding world. From this point of view there is an intrinsic connection between theory and experiment, as there is between science and technique… However, when I take concrete human beings as methodical starting points, I also take as primary human beings’ capacity to conceptualise and evaluate reality. Access to the real aspects of the natural world is made possible by the connection between, on the one hand, our doing (the experimental practices that constitute our operational relation with the world) and, on the other hand, our awareness of that doing. (2013, 279-280, note 1)

Technical operationalism is fundamental to Buzzoni’s account. It is by no means the only account that someone might combine with Kant’s to create a Kantian theory of thought experiments. Nevertheless I approve of it. I would only add that not all empirical content comes from the connection between our experimental practices and our awareness of our doing those practices. In many cases, what is necessary to connect those two things is an exercise, whether conscious or subconscious, of the imagination.

What I like about this account is the emphasis on the cognitive powers of the agent. I have argued elsewhere that some of our most basic epistemic abilities including the ability to interpret, are experimental (Stuart in press), and I have argued in this dissertation that the connection we make in thought experiments between theoretical structures and their semantic content is also experimental. If these claims are correct, then the content of many of our theoretical structures is established only through experiment. And Buzzoni seems to agree:

the particular content of any empirical thought experiment must be, at least in principle, ultimately reducible to sensation (…that is, to experimentations)… In this sense, a thought experiment would be devoid of empirical meaning (that is, it would not be a thought experiment proper to empirical science) if, in formulating and evaluating it, we did not in principle assume an at least implicit reference to a real experiment… The activity of performing thought experiments would be nothing, not even conceptual activity, if it did not possess, even while it is still in our minds, an intrinsic reference to experience. This is the ultimate reason why all empirical thought experiments must be
thought of as translatable into real ones, and all real experiments as realizations of

This is why thought experiments are indistinguishable from laboratory experiments on the
concrete level of analysis. The principle of empiricism ensures that thought experiments, if they
are to contribute to the content of science, are related to experience and that their conclusions
will be empirically testable. And technical operationalism requires that the entities and events
upon which we experiment in thought have their content given by the active, dialectical-
experimental engagement between agent and world. The combination of these two theses
requires that thought experiments always be translatable into laboratory experiments, and that
laboratory experiments always be conceivable as realizations of thought experiments. For
Buzzoni, this makes the two kinds of experiment indistinguishable.

Fehige objects. He claims that thought experiments are not always physically realizable, even in
principle. For example, we cannot summon Maxwell’s demon into a laboratory (no matter what
they tell you on thermodynamic necromancy websites). Nevertheless Buzzoni’s account escapes
this criticism because Buzzoni only requires that all good thought experiments be tied to
experience. He says thought experiments must be realizable by experiment, but Buzzoni can slip
between experience and experiment because his technical operationalism explains the agent’s
connection to their conceptualized experience as arising necessarily through experiment.
Therefore when Buzzoni claims that all empirical thought experiments must be thought of as
translatable into real ones, he means that for each thought experiment, its semantic content is tied
to experience (via experiments), and its hypothesized outcome is justifiable in the sense that its
relation to physical theory makes its claims testable by empirical experiments, which may or
may not be the precise physical realization of the events and objects manipulated in the
imaginary scenarios of the thought experiment. At least, this is how I interpret Buzzoni.
According to this interpretation, Maxwell’s demon “must be thought of as translatable” into real
experiments not in the sense that we can bring a demon into the universe and convince it to sort
particles, but in the sense that the elements of the thought experiment (PARTICLE, CONTAINER,
VELOCITY, AVERAGE, etc.) are tied to experience via experiments we have already performed or
could perform, and the conclusion of the thought experiment can be validated (if we feel that it is
necessary) by testing to see if the second law of thermodynamics indeed has only statistical
violability. Here is Buzzoni in his own words concerning the sense in which the demon must be physically realizable:

To understand the actual technical-operational foundation of Maxwell’s experiment, we must recall that thought experiments, like real ones, cannot be evaluated independently of their intended goals. In order to understand the meaning of an experiment, we must first of all interpret it, tie it to the theoretical formulation of a question about nature and its laws; that is, we must try to identify its intended goal or, more in general, a goal that it may achieve. A thesis upheld on many occasions in this work is that in real experiments a noetic content’s claim to truth is tested by action, and action in turn draws its meaning from that noetic content; this reciprocity is the most basic trait of experimental knowledge. (2013, 98-99)

This could have been a paragraph from my own analysis of Maxwell’s demon in Chapter 5.

Here is how my dissertation extends Buzzoni’s thesis. I have examined how thought experiments can provide content for theoretical structures, which is one way a thought experiment can play the transcendental role identified by Buzzoni. I have also identified the mechanism that is responsible for this role: the experimental connection of a theoretical structure and some empirical content via an exercise of an agent’s imagination. I also sort out (for a small number of cases) how the two different roles for thought experiments (concrete and transcendental) are played, and how they interact. A performance of Maxwell’s demon provides understanding by increasing the (subjective) empirical content of Maxwell’s theoretical structures. A result of this performance is that Maxwell’s thought experiment, and thought experiments in general “anticipate the results of real experiments and in this way they inductively extend our knowledge” (2013, 96). I explain this in terms of the argumentative application of the result of the demon to scientific theory, which can proceed deductively, inductively, abductively, informally, etc. To do this, I tie Buzzoni’s account to the idea that the performance of a thought experiment is separable epistemologically from its theoretical application (Radder 1996, Häggqvist 1996, De Mey 2003).

One difference between Buzzoni and me is that Buzzoni denies that thought experiments are necessary for providing empirical content to theoretical structures. This is because he believes
that our technical-operational situatedness already does that. That is, he adopts the Quinean solution to the problem of coordination examined in Chapter 6. I rejected this as a solution to the problem of coordination because scientists still need to coordinate models and their targets in practice, even if experience is in general already coordinated at many levels through a technical-operational link via experiment. Buzzoni rejects Kant’s notion of a schema due to Quinean considerations, whereas something like a Kantian schema is fundamental to my account. Our disagreement therefore centers on the make-up of the human mind and the source of empirical semantic content. But in terms of the function of thought experiments in science, we agree.

However, my dissertation does not merely concern a transcendental portrayal of thought experiments, it concerns understanding-oriented thought experiments. And this brings me to the work of Catherine Elgin.

2.2 Catherine Elgin

Catherine Elgin has been arguing for some time that thought experiments produce understanding (Elgin 1993, 2000, 2002, 2004, 2006, 2014). Although I suggested some light criticism of parts of her account earlier, I am mostly in agreement with Elgin regarding the characterization of understanding (see beginning of section 5, last chapter), specifically that it is “largely” a matter of knowing where something fits in and how to use it (1993, 15); that it “comes not through passively absorbing new information, but through incorporating it into a system of thought that is not, as it stands, quite ready to receive it” (2002, 14); that it is different from knowledge, that it is often tacit, non-propositional and performative; that exemplification is one device for increasing understanding; and that tools of understanding are not only pedagogical heuristics but include experiments and fictional novels. Finally, I agree that the way we come to understand something, whether through exemplification or experimental imaginative connection (as above), is not due wholly to some power of the mind. It is constrained by the knowledge and understanding we have of the rules, conventions, shared background knowledge and precedents that guide our epistemological practices (1993, 24).

One objection I have encountered to my characterization of understanding-oriented thought experiments is that it reduces thought experiments to mere examples. The intuition is that famous thought experiments like Maxwell’s demon and Einstein’s elevator must be more than examples
because of the role they play in the history of science. To reply, I borrow another idea from Elgin: examples are not epistemologically shabby. What we do with an example is what matters. An example can be much more than an illustration, and in some ways a physical experiment is nothing more than an example of a claim made by a theory (or an example of a counterexample to such a claim). Elgin discusses the notion that life originated on Earth from inorganic materials like water, methane, ammonia and hydrogen stimulated by a bit of lightning. The famous Miller-Urey experiment serves to give us an example of such an emergence of biological matter (Elgin 2014, 223-226).

More evidence of the nobility of examples comes from thought experiments in moral reasoning. Kant argues that we learn morality through case study, just as we learn to apply concepts in general. After all, what is a moral judgment but the application of a moral concept such as, JUSTICE, FAIRNESS, WRONG-DOING, etc.? Elgin (2014) argues persuasively that for this reason, many fictional novels are thought experiments. In real life we are never sure that a case of altruistic behaviour is genuine. However in a novel, the author can simply tell us what the character’s motives are. This way we can be exposed to perfect examples of morally significant actions, traits, states of affairs, etc., and these are crucial for our moral development (see O’Neill 1986, especially 23-29).

I also agree with Elgin that any epistemology that can explain the attempts of science and philosophy to model the world and produce understanding will also explain art’s attempts to do the same (1993, 14). This may be surprising, given that my discussion has centered on theoretical structures. That is, it might seem unintuitive to explain artistic understanding in terms of filling out the meaning of and learning to work with theoretical structures. Nevertheless there is some reason to think that perhaps an account like the one presented above can be useful in explaining some instances of artistic understanding. Not all the theoretical structures I have been considering have been mathematical or mathematically expressed (for example, Darwin’s). As we move away from physics towards the humanities, it seems clear how thought experiments might link theoretical structures to empirical content in chemistry, biology, psychology, sociology, history, literature studies, and so on. A novel could introduce a theoretical structure (as a principle, model, or theory) and then provide empirical content for it through imaginary cases. We are all some combination of rationality, faith (in something) and passion. The Brothers
Karamazov asks what happens when we increase each of those variables while decreasing the others. Fyodor Karamazov has three sons, one predominately passionate, one intellectual, and one faithful. Dostoyevsky could have answered the psychological question in a few words if we all knew and agreed on what we meant by reason, passion and faith. But since he could not assume any shared antecedent understanding, he had to create something that would help us fill these concepts in with content he wanted to discuss, which he does by charting the life events of three imaginary individuals.

Additionally many poets and artists purposely create abstractions that force us to use our imaginations in order to provide empirical content to their creation. (And they are successful insofar as we enjoy exercising this capacity on their work). A term other than “theoretical structure” might be necessary to capture this type of creation, but it would still be explainable using this account. Some of the objectivity of understanding that I argued for in Chapter 7 might get lost here, but that is again consistent with this thesis, since I have been insisting throughout that only when a certain end is accepted does understanding become objective, and the epistemic end of art or literature appreciation differs to some extent from person to person as what they appreciate (value, enjoy) differs.

Moving on, there are a few important issues concerning Elgin’s characterization of epistemology that I would like to mention. By focusing on scientific understanding, she claims that we can bridge the gap between the contexts of discovery and justification. If tools of understanding give us resources for deciding what hypotheses, stances, or modes of categorization are worth taking seriously, they provide some normative structure to the context of discovery. They indicate that discovering \( p \) would be fruitful, whereas discovering \( q \) would not. So they afford some cognitively grounded incentive for investigating \( p \) rather than \( q \). They tell us which truths are worth having. So they enrich the context of justification as well. They not only yield the information that \( p \) is justified, but also that the categories in which \( p \) is cast, the stance from which \( p \) is framed, the questions to which \( p \) affords an answer, and the questions which the confirmation of \( p \) enable us to raise are epistemologically valuable. (2002, 23)
If this is right, investigating the what and how of understanding is crucial even for an objective epistemology concerned only with knowledge and justification. Elgin reminds us that “because of its narrow focus on the conditions for knowledge, contemporary epistemology cannot say what makes insights interesting or important, thus cannot say what sort of knowledge is worth having and seeking. By broadening our vista to include understanding of all sorts, we will be better equipped to deal with the issue” (1993, 14). I have suggested that we can expand epistemology in this way by applying the tools of linguistic interpretation to non-linguistic elements of scientific practice.

To conclude, this dissertation is continuous with and provides support for several of Elgin’s central theses, and is also consistent with the spirit of Buzzoni’s style of Kantianism, although employed in a project of understanding-oriented epistemology.

3 Final Thoughts

Norton’s is one of the most important and influential accounts on offer concerning thought experiments. This dissertation showed that it fails. Paul Thagard is one of the most important cognitive scientists of the last three decades, and this dissertation showed that his skeptical attack against the evidential use of thought experiments fails. Developing a different reading of the paradox of thought experiments, I argued using case studies that thought experiments increase understanding in science by increasing the empirical content of theoretical structures. By “theoretical structure” I have meant the concepts, models, theories and principles that make up purported instances of scientific knowledge. I deny that these structures are never really devoid of content, and insist that what content they do have is relative to the individual. When a theoretical structure is first encountered or devised by a student or scientist, it is often the case that the content can and must be increased for that individual before they can employ it meaningfully. I argued that Maxwell’s demon, Einstein/Bohr’s clock in the box, Einstein’s elevator, Heisenberg’s microscope and Darwin’s eye all increase empirical content in this way. I discussed how thought experiments do this using studies in social and cognitive science. Conceiving of thought experiments as actions, I provided a transcendental justification for the use of the imagination in science. I discussed some current conceptions of understanding and its relation to knowledge and representation. I argued that the understanding produced by the above
thought experiments is an intermediary step between scientific practice and phenomena. I provided support for the claim that understanding does not require explanation. I argued that intelligibility and fruitfulness are not constitutive of understanding, but signs of it. I argued that understanding can be subjective or objective (depending on epistemic aims), and that in science it is usually objective. Finally, I showed that, in general, understanding cannot be equated with propositional knowledge due at least to the role played by models in understanding.

It is my intention that this thesis will help to explain how thought experiments can generate a specific kind of understanding that is necessary for scientific progress. I also hope it is a useful step towards an understanding-inclusive epistemology that draws together the strengths of philosophy, history, sociology and cognitive science. Such an epistemology is necessary for any comprehensive attempt to explain the epistemic product and practice of science. What I have tried to do is reserve a place for the active imagination of the agent in such an account.
References


Boumans, Marcel. 2009. “Understanding in Economics: Grey Box Models.” In Henk de Regt et al. (eds.).


—— 2009. “Ontological Principles and the Intelligibility of Epistemic Activities.” In Henk de Regt et al. (eds.).


French, Steven. 2003. “A Model-Theoretic Account of Representation (Or, I Don’t Know Much about Art…but I Know It Involves Isomorphism).” Philosophy of Science 70: 1472-1483.


2006 Annual Meeting of the National Association for Research in Science Teaching, San Francisco.


