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

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. I will argue that performing a thought experiment can (and does) trigger the imagination in a way that increases the empirical content of a theoretical structure (proposition, model, concept, etc.) for an agent in a way that is epistemologically relevant. This will have 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, Galileo's falling bodies (and pendulums and inclined planes), 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, Cohen 2004), and there is at least one textbook that aims to introduce 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*. 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, thought experiments need not be as grand as these; 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 thus form an extremely diverse set of mental activities. How can we investigate them? First, we have to recognize that their being easy to perform does not guarantee they will be straightforwardly understandable. Expecting to understand thought experiments due to our long experience using them is like expecting birds to understand aerodynamics because they can fly. And 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 birds have evolved to take advantage of them; they do not need to know how flying works. Laws of inference-making are not like this. Our instincts concerning what we should infer from an imaginary scenario are not as reliable. Relying on our imagination to figure out the behaviour of subatomic particles might be like a bird trying to fly in outer space.

To proceed, therefore, we will 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 to transcend the confines of paradigms, which is necessary if we are to be convinced of a new world-view. Kuhn then argues that by changing world-views, we can learn about the world.

Kuhn's answer to the question of how thought experiments fuel scientific progress has not won 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 central for 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 probably is not a paradox. However, according to more inclusive conceptions, it can be. 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, as we will see.

The puzzle transformed into a paradox when Kuhn's open-ended question led to a debate between James R. Brown and John D. Norton. The outcome of this debate was a dilemma between two options: a world with epistemic magic, and one without. Each writer assumes with Kuhn that thought experiments can play a justificatory role in scientific revolutions, and they both take the scientific record as their main source of information. They disagree about what thought experiments are and what they can do. Brown presents a Platonic theory of thought experiments (starting in 1986) and Norton develops an empirical account that characterizes thought experiments as arguments (starting in 1991). Brown claims that thought experiments occasionally provide direct access to truth about the world, something Norton derides as magical. Both writers take the scientific record as their starting point. Each of these are attempts to resolve Kuhn's puzzle, although only Brown makes the connection to Kuhn explicit. Let us see how this paradox emerges from the puzzle.

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). 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: "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 presents Kuhn's puzzle in the form of a dichotomy, and we finally arrive at the paradox of thought experiments. 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.

It is not just Norton and Brown who see this paradox as the key epistemological issue. *In either the abstract or introductory section*, at least 47 papers published since 1991 explicitly mention some form of the paradox (Aligica and Evans 2009; Arthur 1999; Bishop 1998, 1999; Bokulich 2001; Brendel 2004; Brown 1993a, 2004, 2007a; Butkovic 2007; Camilleri 2014; Chandrasekharan, Nersessian and Subramanian 2013; Clatterbuck 2013; Cooper 2005; Davies 2007; De Baere 2003; De Mey 2003, 2006a; Ducheyne 2006; Fehige 2012, 2013; Francis 1993; Gendler 1998, 2004; Gooding 1992, 1994; Häggqvist 2007, 2009; Hopp 2014; Horowitz and Massey 1991; Humphreys 1993; Irvine 1991; Kujundzic 1998; Laymon 1991; Machery 2011; McAllister 1996; McComb 2013; Moue et al. 2006; Nersessian 1992, 2007; Norton 1991, 1996, 2004a; Pitcha 2011; Schlesinger 1996; Shepard 2008; Urbaniak 2012; and Wilson 1991).

Putting it another way, thirty-five percent of post-1991 English-language philosophical items referenced by the (2014) *Stanford Encyclopedia of Philosophy* entry on thought experiments mention the paradox *in the abstract or introduction*. Interestingly, in 2009, *sixty-nine percent* of the relevant literature mentioned the paradox in the abstract or introduction (according to the references mentioned in the *Stanford* entry). This difference is probably due to the influx of papers written since 2009 on the descriptive psychology of thought experiments. The current percentage would increase further if we extended the search beyond the abstract and introduction of the articles in the literature, and looked at monographs as well (then we could include Brown 1991a; Buzzoni 2008; De Mey 2005, 2006b; Gendler 2000; Georgiou 2007; Häggqvist 1996; Sorensen 1992; and others).

Something very important to recognize is that most of the 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). 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 *one* paradox. In the next section, I will map out the conceptual space of the paradox and show that it admits of several interpretations whose differences are indeed epistemologically relevant. One advantage of doing this is that it provides an easy way to sort out the rapidly expanding literature on thought experiments (which I do in section 5, below), namely, in terms of which version of the paradox an author is addressing. Another interesting result is that there are viable ways to formulate the paradox that no one pursues. After some 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, as 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. For this reason, 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

Figure 1 displays nine ways the outcome of a thought experiment can be considered novel.

One way a thought experiment can be considered novel is when its outcome is surprising. I will call this "psychological" novelty.

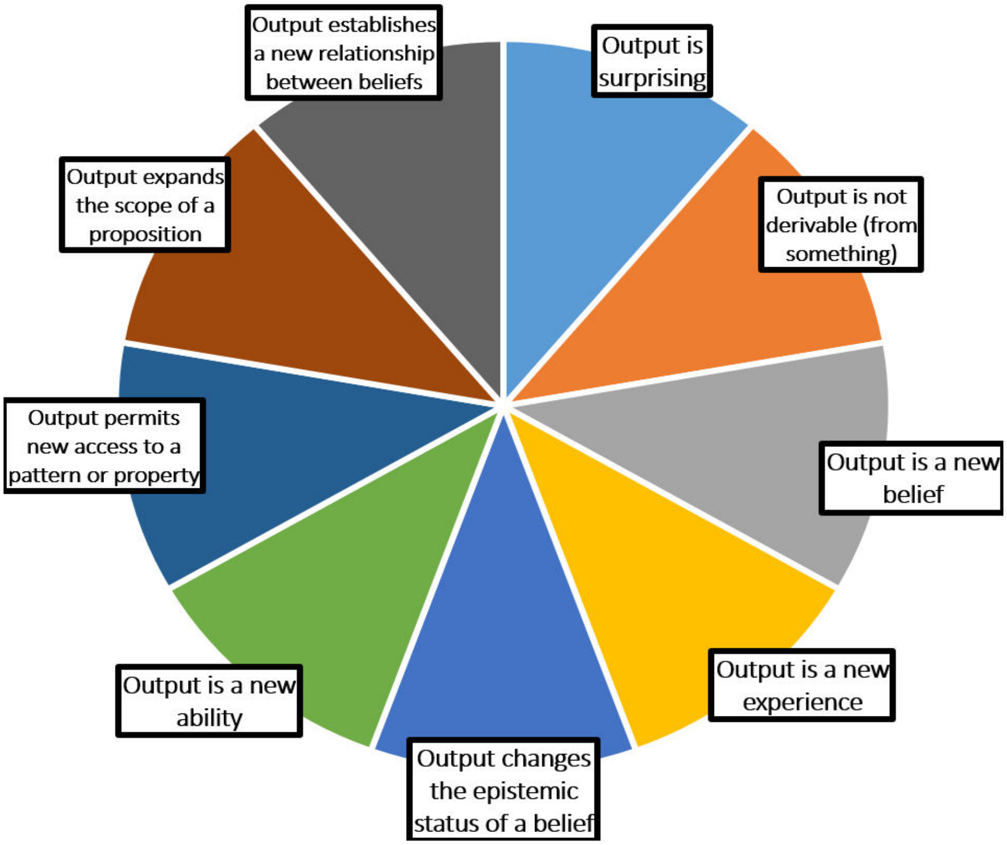
Output is surprising
Output is not derivable (from something)
Output is a new belief
Output is a new experience
Output changes the epistemic status of a belief
Output is a new ability
Output permits new access to a pattern or property
Output expands the scope of a proposition
Output establishes a new relationship between beliefs

Fig. 1: How a Thought Experiment Might Produce Something Novel[1]

[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.

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 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 of a proposition (that we already believed) 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

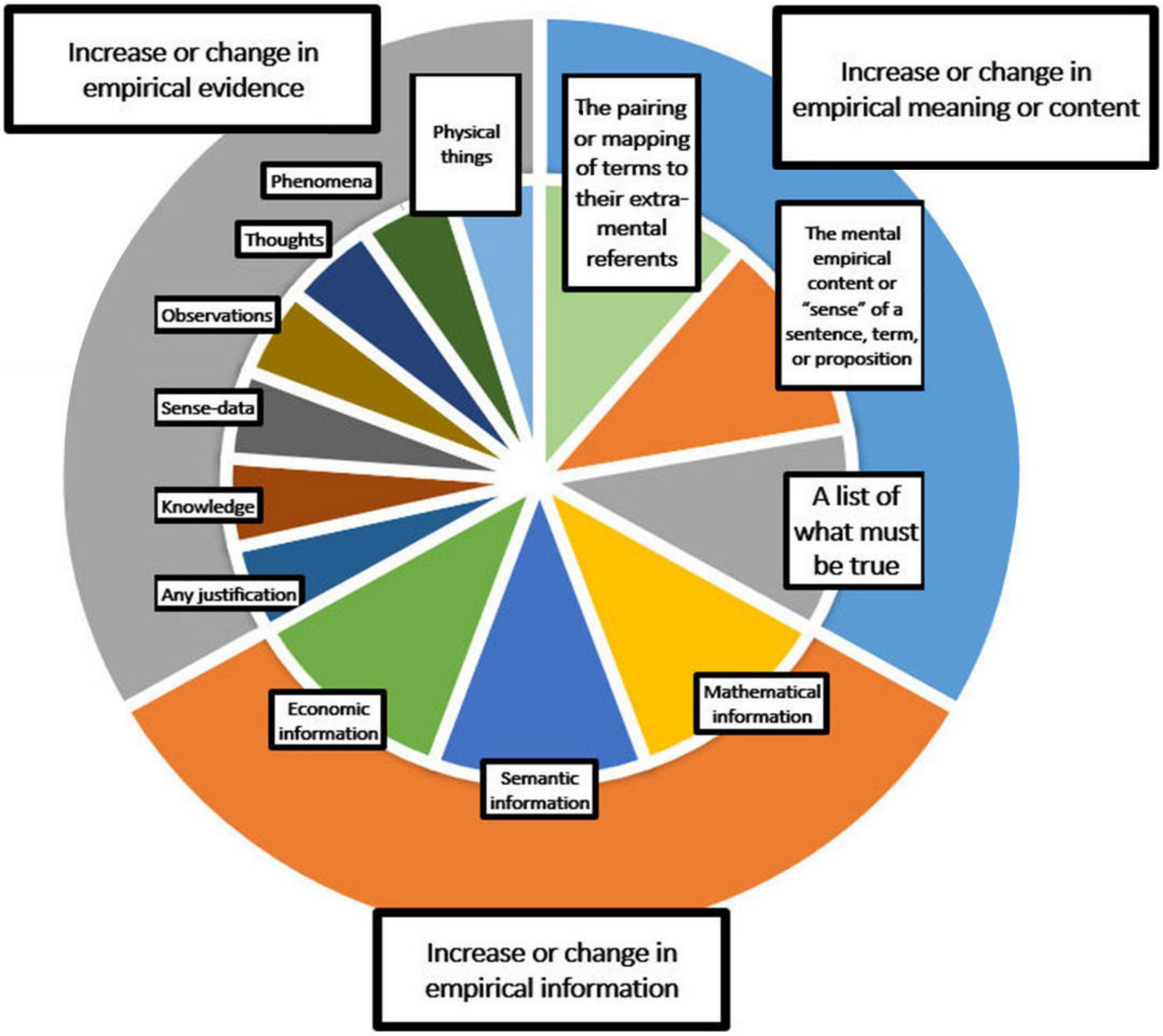
Increase or change in empirical evidence
Physical things
The pairing or mapping of terms to their extra-mental referents
Increase or change in empirical meaning or content
Phenomena
Thoughts
Observations
Sense-data
Knowledge
Any justification
The mental empirical content or "sense" of a sentence, term, or proposition
A list of what must be true
Economic information
Semantic information
Mathematical information
Increase or change in empirical information

Fig. 2: Some Senses of Empirical Import

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 affect (increase or change) the empirical *semantic content* of something, 2) to affect our empirical *information* about something, or 3) to affect our empirical *evidence* for something. As with the different senses of novelty, some philosophers combine several of these senses of empirical import.

3.2.1 *Empirical Import = Empirical (Semantic) Content*

Empirical import understood as empirical content can be interpreted in at least the following three ways. It can 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[2] 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 therefore 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. How should we understand this? 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 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 can be empirical.

[2] In this thesis I use SMALL CAPS to mark concepts.

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 (or others) have experienced. In this sense, ∞ would have less empirical content than BASEBALL, since some of us have seen baseball played, but none of us have ever 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*, even if it has not been experienced. I have never seen the Great Wall of China, but I could. Similarly, I presume that no one has seen a life-sized statue of Shakespeare made entirely of mushy green peas, but such a concept is meaningful because we know what it would be like to experience such a thing.

Let us 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, the flashes must be ignited when the train passes directly in front of her. For the flashes to appear simultaneously to a passenger who is sitting in the middle of the same train, however, the flash at the front of the train must be ignited later than the flash at the rear (from the perspective of the stationary observer), since the light has to travel further when coming from the rear. The conclusion is that the set of events we count as simultaneous depends on our 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 being on a very fast rocket ship and comparing chronologies with 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 semantic content 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 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. This second characterization will be preferable to those who wish to avoid reference to MEANING or INTENSION, which are philosophically difficult concepts.

Each of these interpretations of the way that 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 interesting, these characterizations of the paradox have not yet been explored in depth by those who write on thought experiments.

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 that is completely predictable will only tell us what we already know. Such a message 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 are already certain it will land 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 to 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 *un*informative (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 provides counterintuitive results at the extremes. Given the low probability of most wildly hypothetical statements, these statements will be more informative than verifiable indicative statements (Dretske 1981, 42). “Frank Sinatra is living on the far side of Jupiter as a hamburger” is more informative than “the beer is in the fridge.” While thought experiments do not usually aim to *maximize* information, there *are* often wildly hypothetical. How do we salvage the idea that departures from reality can usefully increase empirical information in the semantic sense?

Floridi avoids the situation where meaningless jumbles of words are maximally informative by limiting semantic information to “well-formed, meaningful, and truthful data” (2011a, 31). By adding in a truthfulness condition, he ensures that we do not allow gibberish to be more meaningful (carry more information) than sensible expressions. According to this understanding of information, then, the paradox of thought experiments would be a question concerning how novel empirical information, that is, novel, well-formed, meaningful, and truthful empirical data, can be produced from entirely within the head. This is an interesting characterization because 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.

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

Now, 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 the properties and uses of specific thought experiments in order to test 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), 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 will not provide "all one's knowledge," they might provide a new piece of knowledge. The same holds for the set of 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 will count as 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 them empirically. Galileo's falling body thought experiment was made empirical, for example, when astronauts dropped a hammer and feather on the moon. Someone might object that many thought experiments are not always empirically confirmable, for example, because they invoke perfectly frictionless planes, demons, or situations where we must decide whether to 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 thought experimental process itself is in the head, or c) the source of the evidence that justifies the output of the thought experiment is in the head.

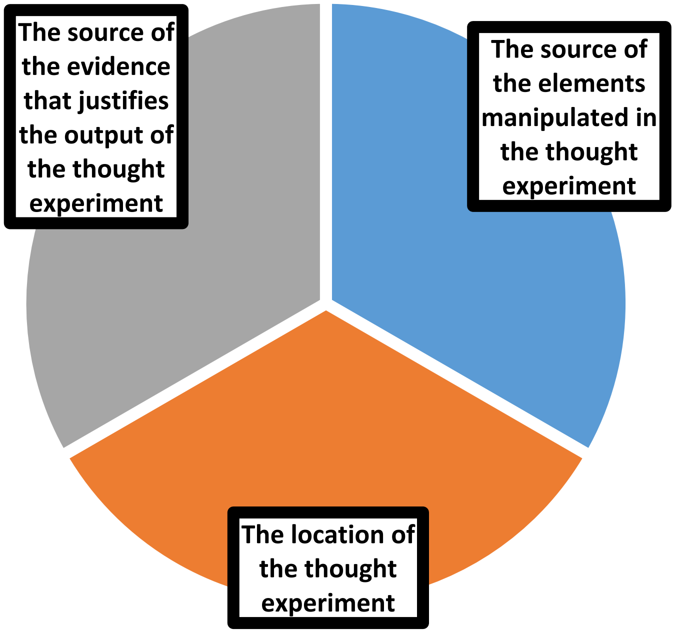
The source of the evidence that justifies the output of the thought experiment
The source of the elements manipulated in the thought experiment
The location of the thought experiment

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 processes of a thought experiment itself provides even more options. Since imagination is involved in most (or all) thought experiments, and imagination is 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 notebooks and smartphones) to count as part of the mind (see, for example, Clark and Chalmers 1998). Some go even further, allowing mental features of others to count as part of our minds (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.[3] 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 complete conceptual space of the paradox can therefore be represented like this:

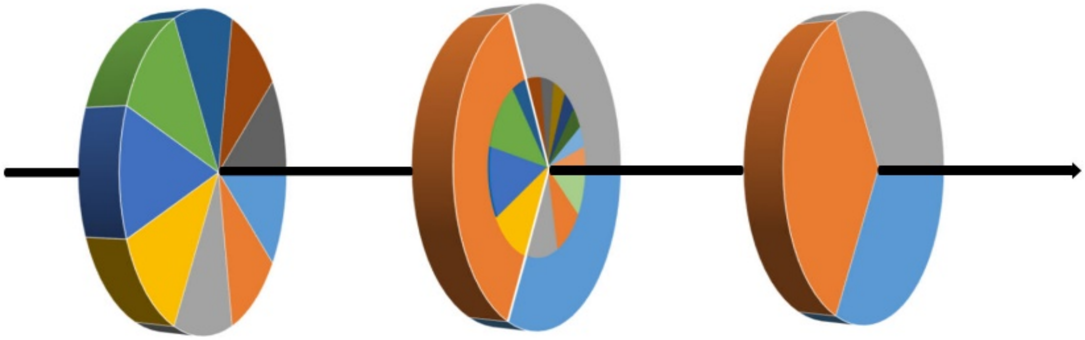

Fig. 4: (Some of) the Conceptual Space of the Paradox

[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.

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 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 for or against those positions. 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 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 all 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 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 should 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, not the other way around.

With all of this in mind, let us see how the main accounts of thought experiments understand the paradox in the different ways listed above, and whether they are 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 takes on new importance because it answers a question we were not asking before. 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 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 counterexamples 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 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. Mediative 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 semantic content. Therefore Brown incorporates at least two of the three main interpretations of empirical import.

Finally, for Brown, the imagined scenarios in a thought experiment are in the head, and not anywhere else. 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 outside of the head. This adds an interesting caveat to Brown's version of the paradox concerning Platonic thought experiments, which asks: how do thought experiments produce (genuinely) novel empirical evidence (in the sense of experience that justifies belief) without new empirical data, *but allowing for new non-empirical data*?

Norton claims a different sense of novelty for thought experimental conclusions. 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 therefore we must explain how thought experiments can justify beliefs that are true. 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 such a 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 means of analyzing two of the 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. Specifically, the mechanisms that 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 that 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 is an important idea that will come up again in Chapters 5-7.

This brings us to the naturalists: those who claim that we can *and should* use 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. This is what makes Sorensen a naturalist in particular: his use of evolutionary psychology to justify the reliability of thought experiments.

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 characterize thought experiments as "mental models," and focus on their performance. 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*. These special skills are usually justified with reference to human evolution.

Let us turn to their characterization of the paradox. 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 perceive what happens in a given imaginary scenario. This is supposed to be very different from working out a solution using logical inferences, as Norton would have it.

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. She claims 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 in order to arrive at 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, everyone agrees that a thought experiment can restructure old information to reveal features of that information we did not notice before. While the structure is new, the content is not, so this falls under the psychological sense of novelty.

However the mental model camp 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, conceiving thought experiments as mental models allows for truly novel mental presentations. But it does not reach Brown, because the presentations are not interactions with real, mind-independent abstract entities. Instead, they are presentations that change the way we interpret our own concepts. For example, Gendler, Nersessian and Miščević agree that rearrangement can help us to "possess" our concepts 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.

Most importantly, these authors also allow that thought experiments can produce genuinely new concepts. When thought experiments do this, they produce something that was not derivable from the propositions given in the set up. Therefore they can produce novel outcomes in the sense that those outcomes were not derivable logically. This is another sense of novelty they share with Brown. However, even if the concepts are new abstract entities, these are not abstract in the Platonic sense.

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. I place Ichikawa and Jarvis here because they fit into the naturalist camp and what they say is clearly consistent with the mental modellers. Of course, it must first be shown 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 display Baconian traits, 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 of the Brown-Norton debate.

For one, the distinction explains why the *initial* response of many students and philosophers is to side with Norton. Given the choice between two Cartesian approaches, one from the assumptions of empiricism and another from the assumptions of Platonism, most people will pick empiricism. However Brown does not employ the Cartesian strategy, and a little 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. Perhaps 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 interesting thing to note 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 counterexamples that would disprove the thesis. This is a common reaction, and several papers have been published which claim to find such counterexamples (for example, Bishop 1999, Brown 2007, Gendler 1998). These papers are trying to fight a Cartesian using Baconian principles, that is, they are looking for *observations* to disprove a fundamental theory. This will not work against Norton. 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 example.

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 convincing inference to be an argument. And this makes his position almost 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 that connects premises to conclusion, where that mark is whatever present or future logics identify as formally reliable. I argue that we should not evaluate thought experiments (or arguments) this way. What makes a thought experiment good or bad is not 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 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 fruitfully (and of course, 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 theories of embodied cognition.

These conclusions are the result of case studies undertaken in Chapter 5, and therefore 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). The conclusions of this thesis are 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 motivate the epistemological account developed 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	Chapter 1
Adopt Norton's Characterization of the Paradox. Show Norton's Account Unsatisfactory	Chapter 2, 3
Adopt Thagard's Characterization of the Paradox. Show Thagard's Account Unsatisfactory	Chapter 4
Identify a New Characterization of the Paradox by Historical Case Study	Chapter 5

Explore the New Characterization using Social and Cognitive Science	Chapter 6
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)	Chapter 7

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. I argue in several places that empirically relevant thought experiments (or arguments, for that matter) will always depend on semi-experimental interaction with experience. Chapter 4 grants that thought experiments can 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. However, I also show that the mental models framework is not enough for a complete epistemological account of the thought experiments I consider in Chapter 5. So in Chapters 6 and 7, I provide a new 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.