A New Function for Thought Experiments in Science

by

Jennifer Lesley Whyte

B.Sc Hons. in Physics and Philosophy, University of Toronto, 2016

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

Of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

Kenneth P. Dietrich School of Arts and Sciences

This dissertation was presented

by

Jennifer Lesley Whyte

It was defended on

July 27, 2023

And approved by

Sandra Mitchell, Distinguished Professor, History and Philosophy of Science

James Robert Brown, Professor Emeritus, Philosophy at the University of Toronto

Marian Gilton, Assistant Professor, History and Philosophy of Science

Dissertation Director: John Norton, Distinguished Professor, History and Philosophy of Science

A New Function for Thought Experiments in Science

Jennifer Lesley Whyte, PhD

University of Pittsburgh, 2023

In this dissertation I propose and defend a new account of thought experiments in science and show that it solves an otherwise outstanding problem in the epistemology of models in science. In the first chapter, I argue that a handful of reasonable premises about the epistemic status of science and its models leads to a challenge: shifts in scientific concepts lead to shifts in scientific models that lead to potential non-empirical incompatibilities between them. The solution I propose is to construe the role of thought experiments in science as non-empirical operational tests of models in a hypothetical context of use – as model engineering, rather than a source of evidence. In the second chapter, I fully elaborate this account, demonstrate its features, and compare it to three of the most prominent alternative accounts of thought experiments within the literature. The final two chapters of this dissertation are case studies that use the model-engineering account of thought experiments to interpret thought experiments drawn from the history of physics. In the third chapter, I present the lottery thought experiment from Ludwig Boltzmann's 1877 paper 'On the Relationship Between the Second Fundamental Theorem of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium' and show that my account not only well-explains the case, but also explains the absence of this thought experiment from the many subsequent presentations of Boltzmann's achievement in this paper. In the fourth chapter I present the Rota Aristotelica, a pseudo-Aristotelian mechanical paradox, and through it discuss the intersection of three topics: thought experiments, paradoxes, and historical variability. I show that my account of thought experiments allows that many paradoxes can be interpreted as thought experiments, and that this way of

interpreting them can solve outstanding questions about what it means to be the solution of a paradox.

My aim in this dissertation is to present a complete picture of an account of thought experiments in science, the way that account fits into contemporary discussions of the epistemology of models in science, and how the account can be used to bring light to historical case studies.

Table of Contents

List of Figures

Preface

This dissertation, and I as its author, owe much to many. I would like to take a moment here to thank all the members of my committee for their ministrations, especially John Norton for his expert stage-managing of this strange production. I am also very grateful to Jim Lennox, Paolo Palmieri, Hasok Chang, Holly Andersen, and all the members of Sandy Mitchell's Idealization, Narrative, and Other Activities reading group for feedback on the various parts of this document across the years of its gestation. A special thanks is also owed to Jim Brown, who got me into this subject, and into philosophy of science in general. This is all your fault.

I would also be remiss if I did not mention my many wonderful friends and colleagues in and around the Pitt HPS community who helped me bring this project to fruition. To the attendees of each WIP where I waved my arms around too much, thank you. I'd especially like to mention the people who, at some point or another, were generous enough to read and discuss this strange beast, like Kelli Barr, Mel Andrews, Caitlin Mace, Dejan Makovec, Alexandra Quack, and Katie Creel. I have found an incredible little world here in Pitt HPS, and without all the trivia nights and Feierabends, without the anagrams on blackboards, the elaborate high-stakes gift exchange, the metaphorical dungeons and metaphysical dragons, I don't know that I could have done it. Most of all, I'd like to thank John Buchanan, the keystone of the arch that supported this dissertation, without whose wisdom and kindness none of this could have possibly occurred.

Introduction

> "Allow me therefore the customary liberty of introducing some of our human fancies, for indeed we may so call them in comparison with supernatural truth which furnishes the one true and safe recourse for decision in our discussions and which is an infallible guide in the dark and dubious paths of thought."
>
> - Galileo, The Two New Sciences, 77. (Galilei 1914)

Let me tell you a story. Once upon a time a man walked up to the top of a tower and dropped a musketball and a cannonball at the same time. An ancient sage, long since dead, had written that the musketball must fall slower than the cannonball because heavier things must fall faster, but our hero was not convinced. So, he devised a clever scheme to trick the ancient sage. If the lighter musketball had to fall slower than the cannonball, what would happen if the man joined the two together? Would the smaller ball act like a parachute, slowing the motion of the larger? Or would the combined weight of the two balls together make both fall faster than the heavier one did before? The ancient sage had no answer, and our hero, who was named Galileo, won the day. The end.

For a discipline primarily concerned with learning about what is, science is suspiciously full of things that are not. From idealized perfect vacuums to abstract infinite lotteries to metaphorical tennis players, science thrives on the imaginary and the impossible. The experimental work of

science might be done in a laboratory with real materials, but the conceptual work is done through impossible stories written on paper, in computers, and in the mind.

The subjects of this dissertation are these fairy stories in science. I will discuss both the objects of scientific fairy stories – scientific concepts and representations – and the distinctive method that scientists use to manipulate them: thought experiments. If real phenomena are the characters of the histories scientists reveal in experiments, then scientific representations are the characters of the fantasies scientists build in thought experiments. A fantasy is not a history, of course, but fantasy is not thereby useless. I will defend an account of thought experiments in science in which thought experiments are the engines of change for scientific representations - where they serve as ways of testing the compatibility and limits of the ways in which scientific concepts are operationalized. To put the moral of this story in the form of a slogan, previous accounts of thought experiments have characterized them as experiments *in* thought, but my account characterizes them as experiments *on* thought. Through a series of whimsical historical vignettes, I hope to show that this way of characterizing thought experiments provides a useful and powerful way of understanding the role those imaginative diversions can play in the history of science.

This dissertation is divided into four chapters.

In the first, I present a problem that science has to solve in order to retain its coherence. A few commonsensical observations about the history of and practice of science gives rise to a justificatory problem for our scientific models. Science does solve this problem, I claim, and one of the ways it does so is through a process that I describe and then identify with 'thought experiments': non-empirical operational 'stress tests' of model engineering.

In the second chapter I put my account into the context of previous attempts to understand thought experiments in science. I show that my account survives a pair of challenges, Kuhn's

Problem and the Ogopogo Problem, and argue that the three main primary competitor accounts of thought experiments in the literature do not. The reason that my account is not affected by these challenges is that my account of thought experiments is non-evidential – that is, it does not presume that the function of thought experiments in science is to provide evidence for or against claims about the world. Most accounts of thought experiments either deny that thought experiments have any role to play in science or construe the role they play as evidential. My account construes thought experiments as non-evidential, but still vital to the project of science. The first two chapters form a pair, one providing the positive argument for the account I espouse, and the other a negative argument against its competitors.

The next two chapters flesh out some of the consequences of the account I provide with concrete case studies.

In the third chapter I show how thought experiments can be used to construct and test new representations. Every method of mathematically representing a worldly phenomenon is simpler than the world it represents. This simplicity is the source of a representation's power, but it also constrains that representation's domain. Finding a balance between grasp and constraint requires non-empirical testing of the kind I discussed in the first chapter, and through two toy examples from applied mathematics and one long case study from statistical mechanics, I show that thought experiments can be used to test variations on that balance.

In the last chapter I turn my attention to negative thought experiments, and in particular, to paradoxes. I explore two broad themes in this chapter: first, the historical situatedness of thought experiments; second, the way that the central tension of a thought experiment can move and change as the conceptual landscape shifts around it. I explore both themes through the long and twisted history of the Aristotelian Paradox of the Wheel, or *Rota Aristotelica*, in three of its most notable

guises. Each of these three ways of construing the Rota Aristotelica arise in a different conceptual context, and the two that survive to the present day survive because they are differentiated in their goals.

1.0. On Fairy Stories (in Science)

> "Stories that are actually concerned primarily with "fairies," that is with creatures that might also in modern English be called "elves," are relatively rare, and as a rule not very interesting. Most good "fairy-stories" are about the adventures of *men* in the Perilous Realm or upon its shadowy marches. Naturally so; for if elves are true, and really exist independently of our tales about them, then this also is certainly true: elves are not primarily concerned with us, nor we with them."
>
> - J.R.R. Tolkien, "On Fairy Stories" (Tolkien 1966)

1.1 The Sleep of Reason Produces Monsters

I am going to point out something very obvious and then show that the obvious point I just made has enormous implications for philosophy of science.

Models in science require non-empirical tests as well as empirical ones.

This requirement is analogous to the necessity of checking a system of propositions for logical coherence. However, the nature of scientific models and the way in which they are used by science mean that the process of checking their coherence is both more complicated and more contextual than that of their propositional counterparts. That process of checking, which I shall characterize as thought experiment, will be the focus of this dissertation.

First, I will dispense with intuition and more firmly establish the truth of the so-called obvious claim I made above. I believe this claim falls out of a reasonable and commonly-held set of premises about scientific change and scientific models. The first premise is that science is in a state of constant change. The second is that the majority of scientific change is change in scientific models and how they are used. The third is that scientific models get their goodness from their usefulness, and that usefulness is defined only in the context of use with other models. Those three premises together, I argue, lead to a *prima facie* problem for the justification of scientific claims. Nearly all our claims are made using models whose goodness depends on a huge and constantly-shifting morass of other models, none of which have any a priori reason to not conflict with each other. Once the challenge of non-empirical tests for models is on the table, I will present an answer: thought experiment. Thought experiments, as I characterize them, are one of the ways that science non-empirically tests the coherence of its models. Construing thought experiments as an answer to my challenge requires a new account of thought experiments. This chapter and the next form a pair. I will present the positive case for that new account of thought experiments, the Model Engineering account, in this chapter. In the next, I will make the negative case that the Model Engineering account is not vulnerable to problems that other accounts of thought experiments experience.

1.1.1 Constant Change is Here to Stay

Science changes over time. That seems like quite an obvious observation but taking it adequately seriously at the outset of an account of the conceptual structure of science requires certain sacrifices in how we consider the role of truth within science. In particular, it forces us to confront the fallibility and changeability of the way we currently conceive of even the most fundamental scientific entities. The reality of scientific change does not undermine the success of science, but it does force us to uncouple that success from the unchanging capital-T Truth.

Upon its release in 1962, Thomas Kuhn's *Structure of Scientific Revolutions* forced philosophers of science into the task of explaining historical scientific change as radical conceptual change, and not just incremental progress towards truth. Kuhn thought of scientific change as proceeding in a sequence of puzzles and solutions. Normal science consists of an orderly and unproblematic puzzle-solving activity whereby a scientific paradigm[1] generates problems that need to be solved within it. Each successful solved problem spawns new problems. However, a puzzle devastating to the internal coherence of a particular conceptual scheme could arise from it and necessitate a scientific revolution. These anomalies are unsolvable puzzles so severe that they throw the whole conceptual scheme into disarray and meaninglessness. The fault they reveal is so devastating that it becomes impossible for those that recognize it to continue to think along the lines of the given conceptual scheme. Kuhn's primary examples were dramatic - Galileo's devastating attack on the Ptolemaic world system in the Two Chief World Systems, for one. The dominant way of viewing the cosmos seemed to change overnight, and all the centuries of careful mathematical work on epicycles and deferents that came before were put by the way (Kuhn 2012).

However, in the years following Kuhn's breakthrough other philosophers attempted to somewhat dampen the drama of Kuhn's account. After all, few scientific episodes are as dramatic as the Copernican Revolution, and most scientific change happens within the period of Normal Science. Imre Lakatos proposed an account of scientific conceptual change that kept the central insight of Kuhn's approach but accounted for the subtler changes as well. Lakatos thought of a scientific conceptual scheme and its attendant practices (a 'research programme', in his terminology) as composed of a large set of theoretical claims, some more central than others. If the central claims

[1] For the purposes of this section, I am interpreting Kuhn's 'paradigm' in the broad sense, rather than the narrow. In his later work, Kuhn favoured the narrower sense, in which a 'paradigm shift' is a much smaller and less dramatic event. However, the broad sense, with all its drama, is still clearly present in Structure, and has hence been extremely influential.

were threatened, new theories could be taken on board as armour. These auxiliary theories support the central claims by restricting the set of possible counterexamples to those central claims. A research program can add and jettison these auxiliary theories as necessary, and principle could do so forever. However, a research program that relied too hard on the introduction of new auxiliary hypotheses to protect its central claims would eventually lose the ability to solve new problems as they arose. The plethora of new claims weaken the descriptive power of the central theory by allowing more and more exceptions to those central claims to be permissible under their aegis. As Lakatos learned from Popper, a theory that explains everything really explains nothing. When a research program begins to degenerate, its leaner competitors may eclipse it. On Lakatos' view science is constantly in a kind of research and development arms race. Different scientific research programs attempt to build the most effective and efficient science machines they can - the system of claims least vulnerable to attack from their enemies and most able to handle the new challenges that come their way (Lakatos 1978).

So, between these two authors there are two kinds of scientific change on the table: the dramatic paradigm shifts of a Kuhnian Revolution, and the gradual build-up of conceptual content that characterizes a Lakatosian programme. These two kinds of change are now taken as given within philosophy of science. It is consensus amongst philosophers of science that the activity of science produces conceptual change of one of these kinds. Were science not revising its concepts, that would mean scientists were not trying to use their concepts in new ways, not looking for new challenges from new phenomena, and not attempting to solve the trickier puzzles – in short, they would not be doing science. In Section 2 I will describe in greater specificity how I think the conceptual change in science falls out of its status as an activity, but for now, it will suffice to note that a science without conceptual change would not be a science we should esteem. If science is worth doing, it is because it has the potential to produce change in how we think about the world. If

we were certain that our conceptualization of the world around this was perfect and unchangeable, the great labour of science would be in vain. Constant change is here to stay.

1.1.2: Models are the basic unit of scientific change

One of the great innovations in philosophy of science in the latter half of the 20th century was the move away from describing the activity of science as being primarily manipulations of either formal propositions about a given natural phenomenon or the natural phenomenon itself[2]. In truth, most of the practice of science is neither so abstract nor so concrete. Most of the activities of a working scientists are manipulations of scientific models – intermediaries and mediators between description and described. Models are not descriptions of scientific phenomena, but objects of scientific study in their own right. They are typically idealized, abstracted, or lightly fictionalized[3] representations of the subjects of scientific inquiry. They are often (but not necessarily) mathematical in nature. And, definitionally and crucially, models serve as surrogates in reasoning about the system in question. This is the distinctive power of a model-based approach to philosophy of science: it can justify why meteorologists work with computers rather than clouds, why chemists are taught with ball-and-stick toys and not acids, and why the average theoretical physicist need not personally boot up the Large Hadron Collider every morning in order to do their day's work.

[2] The prime mover of this shift is Mary Hesse's 1963 book *Models and Analogies in Science*, which set the terms of the debate for the remainder of the 20th century (Hesse 1966). The volume of literature that has sprung from this central insight is too vast to summarize. The approach I take here is heavily inspired by Mary Morgan and Margaret Morrison's discussion of models as mediators (Morgan and Morrison 1999).

[3] For the purposes of the account presented here, the fine distinctions between idealization, abstraction, and fiction are not relevant. Moreover, there is limited consensus on how to draw those distinctions or on what features of those possible distinctions is epistemically relevant. I will not depend on any fine-grained notion of idealization, abstraction, or fiction in what follows – I require only the extremely minimal claim that models are less complicated depictions of their targets.

Both scientific theories and models are constructed from scientific concepts, and in turn, theories and models give scientific concepts their content. A scientific concept is a complicated entity with many parts – the concept 'mass' features in everything from casual statements about how much produce you can get for your money, to the 'm' in F=ma, to more arcane derived concepts like 'rest mass'. 'Mass' is a particularly successful concept, so it finds its way into many different theories and a vast panoply of models. Less successful concepts are confined to more limited domains. A concept that only has a single context of use is typically of little interest. Proprietary concepts cannot aid in the broadening of the understanding, in the same way that even a perfect understanding of the chess queen cannot help a player understand real monarchies. However, a consequence of all interesting scientific concepts having many different domains of application within different theories and models is that the addition of a new domain to a concept or a change within any of those domains constitutes a change in the concept writ large. Copernicus' extension of the concept 'planet' to include the Earth changed what it meant for Mars to be a planet, just as the IAU's later splitting of the concept into 'planet' and 'dwarf planet' did hundreds of years later. If the changes cause the concept to become incoherent, the concept may split, or its users may bar a deviant use[4]. Regardless of what the concept's ultimate fate is, the whole concept is implicated in its transformations.

If science is a war, theories are the generals and models the foot soldiers. A few charismatic and well-known theories command a vast host of models, each of which does a small part of the work of science. As many authors[5] have pointed out, the relationship between the generals and the

[4] There will be more discussion of the barring of these monstrous uses later in this chapter. For now, though, I invite the reader to consider cases of negation-by-further-clarification, such as 'Vegan Leather', 'Potemkin Village', and, indeed as it is conventionally used, 'thought experiment'.

[5] For instance, Sandra Mitchell argues that as soon as philosophy expands its analysis of science beyond physics, any simple hierarchical analysis of theory falls away immediately (Mitchell 2009). On the other hand, Nancy Cartwright has argued that even within physics the relationships between theoretical laws are deeply complex and context-dependent (Cartwright 1983).

foot soldiers can be very complex, and the precise hierarchy varies by division and mission. The models deployed in quantum interpretation are closely tied to the theories they serve, whereas the models deployed in the social sciences are often quite situation-specific and distantly removed from the headline theories of their research programmes. The broad aegis of a theory can disguise profound differences between models used within the same discipline. Mark Wilson, for instance, describes theories as mere 'facades' that lend apparent harmony to what is in truth a radically disjoint set of practices within a given science (Wilson 2006). The broad theories a scientist will quote in press interviews and the introductions of textbooks often have very little to do with the work that goes on in quotidian practice, and the quotidian practices of two labs under the same theoretical banner may have very little to do with each other[6]. The engineer building with rigid bodies and the materials scientist constructing new and strange alloys may both be involved in the activity of building a bridge, but that doesn't mean that any two of their models will be even remotely similar to each other.

How does this diversity come about? It falls out of the heuristics of modeling. When a putative modeler sets out to represent the intriguing new phenomenon on her laboratory bench her first port-of-call will always be to models of similar phenomena. The mathematical sciences have a vast canon of modeling methods that work in particular areas, and there's no sense in reinventing the wave equation. George Polya's evocative recommendations for the mathematician faced with a new problem provide an excellent encapsulation of the practice of scientific modeling as well. In *Mathematics and Plausible Reasoning* Polya takes the reader through a series of solutions to tricky-seeming problems that he arrives at in each case by modeling the problem as some other analogous-

[6] For Wilson this diversity (particularly diversity between different scales of the same phenomenon) is vital to the descriptive adequacy of science, since different models need to 'borrow' facts from each other in order to properly set the boundaries of their own applications. See (Wilson 2017) for further discussion of this point.

seeming mathematical structure (Polya 1954b). This process is carried out largely in the absence of theoretical considerations. Apparent analogy can sometimes be theoretically explained later, but not always. The same is true in science. Indeed – for data-intensive sciences in which modeling typically takes the form of fitting an equation to a dataset, the process of modelling is so independent from theory that it can be automated by an algorithm that will test various typical models until it finds the best fit. In addition, as Patrick Suppes notably noted, even the process of data collection itself is done with a model of the data in hand, that shapes the data into a comprehensible prior to any attempt at understanding how that shape fits into a bigger theoretical picture (Suppes 1962). Modeling precedes theory. Working scientists use whatever models are at hand and change them as necessary.

The distinction between the general, abstract theoretical level and the many levels of models between it and the phenomena allows us to account for both kinds of scientific conceptual change discussed above. Theories change slowly and often dramatically, and a theoretical change typically requires a large-scale conceptual revision, like the paradigmatic Kuhnian revolution. However, models change whenever they are used for a new purpose, often in an ad-hoc way, and the conceptual change induced by these changes in use are typically more subtle, like the slow spread of a Lakatosian change. As I have argued above, every change to a use of a scientific concept changes the whole concept. The implication of these two points is that the *vast majority* of scientific conceptual change results from the attempt to use and extend scientific models, rather than reasoned debate over theories. The changes are smaller – typically just little added contexts around the fringes of a concept – but they are much more frequent. The main driver of scientific conceptual change is not a grand theoretical revolution every century, but a thousand little conceptual extensions every day.

1.1.3 Patchworking

The turn towards a focus on models in science also uncovered new epistemic problems lurking just below the surface, and old ones in new guises. There are two fundamental questions that any account of models must answer[7]: First, the model and the phenomenon it models are related by a 'representation' relation – what does this relation consist in? Second, some representations are clearly better than others for our scientific purposes – how do we explain this? A third question, though one that has been much less discussed by the literature is this: do we know whether or not our models stand in the appropriate relation to their target phenomena, given our ignorance of the target phenomena? If so, how do we know? Models are, minimally, tools for inference about a particular target phenomenon. Whether they are more than mere tools is an open and hotly contested question.

A good account of representation will provide satisfying answers to all these questions. An excellent account will have satisfying answers to these questions that provide normative guidance for real science. Though there are many different accounts of scientific representation in the literature, they mostly break down into two large families delimited by how they answer the above questions.

Mapping accounts characterize the representation relation as a map of some kind that obtains between the putative representation and its target. These accounts are an evolution of older 'similarity' accounts, which proposed that one system represents another if it the former and the latter are similar in some respect. This criterion was too imprecise to do the job. As Nelson Goodman pointed out in his influential critique of similarity accounts of representation in art, all things are similar to each other in some respects, and dissimilar in others – thus, a representation relation defined in terms of similarity alone is vacuously true of all pairs of things (Goodman 1968).

[7] These first two questions, and this framing of the subject, are adapted from (Suárez 2003).

More sophisticated mapping accounts, therefore, typically give a more specific and concrete account of the relation that has to obtain between representer and represented. Different mapping accounts propose different maps – partial isomorphisms, cashed out in a variety of ways, is a typical candidate, as is homomorphism[8]. When the appropriate mapping is present between the systems of interest, the representation relation obtains. The presence of the mapping is also the explanation for the success or failure of surrogative reasoning performed using the model. If all the relevant parts of the target system are present in the representation, then any inference performed with those parts should hold for both. The epistemic question, however, is a persistent problem for such accounts, since the true structure of the world is not known to the researcher who is representing it. Except in artificial situations, we are not in a position to actually claim that any success attained by a given model is due to the presence of a mapping from theory to world. We can say that structural similarity in artificial contexts where we have access to the 'true structure' of both representation and represented can allow the activities undertaken using the representation to transfer over to that which is represented[9]. However, we are never in a position to make this claim when the thing that we are representing is the natural world. We simply do not know what the true structure of the world is. That ignorance does not seem to affect our ability to reason about the world. So, even if some similarity with the world is necessary for successful surrogative reasoning, knowledge of the mapping is not necessary. In addition, the presence or absence of similar structures in the representation and represented isn't sufficient to guarantee the *usefulness* of a model for a given inferential purpose. A representation that stood in a perfect one-to-one map of its target, for instance, would be useless for the purpose of surrogative reasoning, since it would be no easier to

[8] There are many different ways of cashing out the content of the representational mapping. See (Frigg and Nguyen 2021) for a comprehensive account of these subtly different proposals.
[9] For instance, it is the structural similarity between the meter of 'A Whiter Shade of Pale' and 'The Muppet Show Theme' that allows each to be sung to the tune of the other.

reason with than the world it represented. The partiality of a representation is often precisely the point of using one. These critiques, put together, undermine the adequacy of the mapping account in two ways. We cannot know that mappings were responsible for past successes in surrogative reasoning. We also know that the presence of a good mapping does not guarantee the success of surrogative reasoning, since a perfect map is useless. Thus, the presence of a mapping is not the thing that makes that representation good for actual scientific purposes. The thing that makes the difference is whether the similarity allows for surrogative reasoning, and that is not reduceable to the presence or absence of a mapping.

The main rival to mapping accounts of representation are the various inferentialist theories of representation. Inferentialist accounts take surrogative reasoning itself to be the defining feature of scientific representations. These accounts typically deflate the representation relation itself to be merely stipulative, following (Callender and Cohen 2005). A purely stipulative account of representation affords the representation relation itself very little content. If I say 'let's take this dodecahedron to represent the celestial aether', this dodecahedron now represents the celestial aether whether well or poorly. This deflation means that the representation relation itself doesn't carry much epistemic weight. Inferentialist accounts thus must re-inflate the account by identifying the goodness of a representation with its inferential utility: a representation is good iff it supports good inferences about its target subject matter (which the aforementioned representation of the solar system via shelf dice presumably does not). There are good reasons to favour an inferentialist account of representation over a mapping account. The inferentialist accounts paint a picture of scientific representations in which representations are not pictures being revealed, but tools being made. It has no explanation to give of why a particular representation is more useful than another.

I intend my account, presented below, to go some way towards filling this gap in the inferentialist account by describing part of the process by which models are improved and checked. Science makes its conceptual decisions for reasons, so looking at those reasons should show us what makes models better for the actual scientists who are using them. I will assume a loosely inferentialist account of representation in what follows.

Inferentialism, though, has a consequence that a mapping account does not. An inferentialist account of models does not provide us with any hope that some sunny day we might discover the final model that represents everything perfectly. A mapping account offers us the seductive spectre of the True Structure of reality, captured within our mathematical grasp. An inferentialist account pushes us instead towards building the best models we can in particular circumstances, and judges those models by how well they work in those circumstances. It may simply be the case that the universe is not amenable to unified description, and if that's true, then the inferentialist must accept that the best possible models will still be a piecemeal motley.

It is certainly the case that our current slate of scientific models is a piecemeal motley. As William Wimsatt argues, the idea of a science composed only of beautiful and unified truth is an idea of a science for gods, not limited beings like ourselves (Wimsatt 2007). Human science is built out of millennia of rules of thumb, heuristics, kludges, and patches. A well-trained scientist not only knows the models of their discipline, but also how to apply them carefully in only the domain for which they are useful, tiptoeing around the regions in which it doesn't work and the unruly limits that result from stretching its parameters too far[10]. No science is free from this situational boundary hedging – it is as present in the Effective Field Theories of physics as the huge constellation of

[10] The oft-tacit knowledge of how to apply models within appropriate domains and not others are what allow sciences to truck along even with apparent descriptive contradictions. One context in which these apparent contradictions are common is when scientists must reason about phenomena across different scales. See (Batterman 2013) and (Wilson 2017) for discussions of how scientists navigate this descriptive complexity in the context of scales.

models that cognitive scientists wrangle into experiments. This plurality is typically the kind that Sandra Mitchell calls 'compatible pluralism'. Compatible pluralism, in comparison to 'competitive pluralism', is a model of science in which many different descriptive strategies are used in tandem despite the *prima facie* contradictions between the different descriptions (Mitchell 2002). An unskilled practitioner could mistake the complexity for contradiction, but in practice all the participants can navigate the varied ground well enough to not trip up. As Nancy Cartwright notably argued, the world rarely, if ever, presents us with one cause operating at a time, and as soon as there are multiple causes in a single context, we require multiple models to capture them (Cartwright 1983). The complexity of the subject matter of science itself, then, guarantees the multiplicity of its models. There is no model that truly works alone – each must be used in tandem with its near neighbours in order to describe anything beyond the very narrowest of contexts.

When this insight is combined with inferentialism about representation, it produces a notable result: the goodness of a given model can only be appraised in the context of all the surrounding models with which a given model is used. If the goodness of the model depends only on its inferential utility, and a model can only be useful in the context of its brethren, then it is never an individual model that is good or bad, but a model in a particular context of other models[11]. Models are buried deep in their contexts, and we cannot hope to free them. So, we must deal with models within their contexts. We cannot sew our science from whole new pieces – we must use patchwork.

[11] And the context of a model includes all aspects of the use of the model: who is using it, for what purpose, in what ways, at what times. Giere claims that the representation relation itself has to include all of these aspects. I do not hold that all these aspects are contained within the representation relation itself (since I hold a stipulationist view) but it's clear that these contextual aspects are highly relevant for the *success* of representations.

1.1.4 Shifting Sands

I have argued in the three previous sections for three theses: first, that science changes constantly; second, that the basic unit of scientific change is the model, not the theory; and third, that models can only be appraised in the context of use alongside other models. These three theses together pose a problem for science. If the goodness of our models, the foot soldiers of our science, is only meaningful in their context of use and alongside other models, and all those models are constantly changing, then we won't know whether our models are going to work until we try to use them. Were the basic units of scientific practice theories, instead of models, we could check them against each other for logical consistency in the manner of any philosophy undergrad. But models need not be logically consistent, and they can be used in tandem with other models with which they are not logically consistent, as we have established, so long as that logical inconsistency is outside the relevant domain. How, then, do we judge whether our models are any good? To make the problem sharp, let me introduce a distinction.

In order to be useful *qua* model, a structure must be both well-fitted to its target and internally coherent in the right kind of way. A model can fail to be useful for either reason. The question of what makes a model well-fitted to a particular domain is typically an empirical question, addressed by empirical means – let us call this process *model application*. The familiar practice of curve-fitting is model application. But the question of what makes a model internally coherent enough to actually be used for a given function is not an empirical question. It has to be addressed at the level of the model itself, independent of the way the world is. Let us call this process *model engineering*. Consider a pair of leather boots – in order to be good qua boots, the boots must be the right size and shape for the feet in question, but they must also be sealed well enough around the

soles to keep the rain out. When asking if a pair of boots is good[12], one must ask both questions: do the boots fit? And are the boots themselves well-made? In this analogy, the question of whether the boots fit is model application, the question of whether the boots are well-made is model engineering.

These twin qualities need to be tested differently. Model application must be tested by application to the world. Model engineering needs to be tested by simulated use within a proposed context. We cannot hope to test questions of model engineering empirically, any more than we can test whether a set of propositions is logically consistent by trying to derive a proposition from it. Any proposition can be derived from an inconsistent set of statements, and a paradoxical model could similarly match a given data set without thereby representing it well. Questions about the coherence of the model must be addressed at the level of the model itself – the semi-abstract context of use. This simulated use is what I will call a *thought experiment*. Let us investigate how it works.

1.2 The Monster Mash

In order for a model to be shown to be useful, it must be used. For boots to be well-made, they must be able to retain their structural integrity in the context for which they were designed – walking[13]. How do we test a model? In what follows, I will argue that a *thought experiment* is a procedure undertaken in order to hunt for *monsters* within *relevant operational contexts*. First, I will define *monster*, *operational context*, and *relevant*. Then, I will put all the pieces together to show that a thought experiment, so characterized, solves the challenge I presented in the first section.

[12] The failure of Sam Vimes' boots is a failure of boot engineering, not boot application.
[13] It is the purpose for which they were made, so it is the purpose that they shall fulfill. (Sinatra, 1966)

1.2.1 What is a monster?

In Lakatos' second dissertation work, *Proofs and Refutations,* Lakatos presents a dramatized dialogue between an instructor and a group of students arguing over the concept of a polyhedron. It seems like a strange thing to argue over - after all, a polyhedron is as simple a mathematical concept as one could hope to encounter. It is as ancient as the Pythagoreans and as basic as an elementary curriculum. Yet Lakatos identified something strange about the history[14] of the concept of a polyhedron. It did not simply spring out of mathematical practice full formed and obvious - it needed to be formed intentionally. In Lakatos's historical tale, mathematicians are presented with a hypothesis about polyhedra (the 'Eulerian Lemma' that for all polyhedra V – E + F = 2, where V is the number of vertexes, E the number of edges, and F the number of faces) and are asked to attempt to either prove or refute it. The hypothesis is not initially taken to be *definition* of 'polyhedra' but merely something hypothesized of them. The hypothesis holds sound over the simplest and most paradigmatic polyhedra, but as the students grow more ambitious they propose polyhedra that no longer seem to accord with the hypothesis. For instance, some stellated polyhedra do not agree with the hypothesis, and nor do polyhedra that are not convex. The dialogue then becomes a debate over whether or not these abnormal polygons, this stellated and nonconvex polyhedra, should be counted as standard polyhedra at all. Perhaps there is a subclass of polyhedra defined by their ability to fit the initial hypothesis. Or perhaps the hypothesis is merely false – the set of all polyhedra contains the abnormal ones and the hypothesis does not obtain of them.

Lakatos names these strange polyhedra 'monsters'. Monsters are the shadowy entities lurking around the dark edges of our bright and clear theoretical terms. Monsters appear to be legitimate

[14] Or perhaps 'history*'. Though Lakatos drew from real historical mathematical sources for the fictional debate within Proofs and Refutations, the story has been intentionally reconstructed and idealized to fit the philosophical tale Lakatos desired to tell.

members of the domain of application of a concept or model, but their inclusion produces problems for the integrity of the concept or model. In the Lakatosian cases, the monsters scupper the proofs the students are proposing. Though *Proofs and Refutations* deals with only two cases of mathematics encountering and fighting off its theoretical monsters[15], Lakatos clearly intends the moral of the story to be more general. The definition of a concept, even a very familiar concept, does not merely solidify out of its vaguer intuitive meaning: it must be hewn out of vagueness by a series of intentional decisions prompted by the monsters encountered in its context of use. In mathematics, this context of use is the smaller lemmas that attached to a particular defined term (in this case, polyhedron).

Lakatos goes on to explain that conceptual progress in mathematics is often of this character: not the Euclidean definition of terms, exceptionless and pure, and their logical consequences; but a messy and protracted process of finding the exceptions to general-looking principles and deciding, dialectically and dynamically, what to do with those exceptions. Are the monsters to be barred, like a polyhedron with a hole in it? Or are they to be enfolded into the concept of polyhedron and the domain of the lemma restricted instead, as the small stellated dodecahedron eventually was? In Lakatos' picture the decision is not forced upon the mathematical community by definition or precedent. The concept of a polyhedron itself is not rich enough to answer the question the monsters pose – only an active decision on the part of the mathematicians could answer it.

[15] A dearth of examples for which Proofs and Refutations has come under persistent critique. The bite of the critique is that prolongued conceptual debates in mathematics similar to the ones described in the book are very scarce, if not completely nonexistent within the broader history of mathematics. Lakatos himself only ever provided one other example, which was appended to the second edition of Proofs and Refutations as 'Cauchy and the Continuum' (Lakatos, Worrall, and Zahar 1976). Though this critique is sharp if we consider Proofs and Refutations to be making a substantive descriptive historical thesis, it has no bite if we consider it to be largely a philosophical text about the behaviour of concepts. That is the spirit in which I shall take this text going forward.

Though Lakatos' mathematical program is widely read by historians, philosophers, and especially teachers of mathematics, it has perhaps not been as influential within philosophy of science as it deserves. The picture he presents of mathematical concepts is radically different from conventional wisdom about the nature of mathematical meaning. For Lakatos, mathematical concepts exhibit the property that Friedrich Waismann calls 'open texture', though Lakatos does not cite Waismann[16] in this context. A concept is open textured if there is a possible domain of the term's application for which it is unclear whether or not the concept does or does not apply, pending further speculation[17]. Open texture is distinct from vagueness – vagueness, to Waismann, is equivocal use (his example is the word 'pink', which is vague since it refers to a large variety of colours, many of which are very dissimilar to each other, and only debatably applies to any particular case), whereas open-textured terms are precise in their use up to a point. An open-textured term is a term that has a plausible situation of application in which its application would be unclear, regardless of whether or not that situation ever obtains. Waismann's example of choice is 'gold' – gold is not a vague concept, since a given atom or bar of metal or item of costume jewelry will always unequivocally be gold or not, but there are possible scenarios in which it would be impossible to apply the concept 'gold' without arbitrariness. If a putative ingot of gold exhibited a property that gold is not typically taken to exhibit (say, if it glowed faintly, or to use Robert Boyle's much earlier example, if it was made in an alchemical laboratory instead of the bowels of the earth) then it has revealed a bit of open texture within that concept – it is, in Lakatos' terms, a monster.

[16] It is unclear whether Lakatos was aware of Waismann's term. He cites other work by Waismann, but not the works in which Waismann defines and defends the concept of open texture. (Tanswell, Rittberg, and Larvor 2022) have recently shown that Lakatos had some contact with Waismann's work through the mathematician George Kneebone, who was an avid reader of Waismann. However, unless more documentary evidence comes to light we may not assume that Lakatos had Waismann in mind while writing Proofs and Refutations.

[17] Waissman's original presentation of this concept is in (Waismann 1947). However, his definition is slightly obscure. More accessible contemporary definitions are given by (Blackburn 2008) and (Shapiro and Roberts 2021).

For Waismann, most of language is infected with open texture, with the only *a priori* domains like mathematics or rigidly circumscribed domains like a chess game as exceptions. Lakatos' examples show that even mathematics is not safe[18]. The ambiguous edges of our concepts lurk in even formal conceptual domains, and the best we can do is try to make good decisions when we encounter them. What would it mean to allow the alchemist's gold to be gold? Would that improve or worsen our descriptive power over gold[19]? These decisions, if adopted by the broader scientific community, are the minor conceptual changes that allow science to evolve over time outside the context of Kuhnian revolution in the way I described above.

The account I will provide below extends the Lakatosian programme in two ways: First, I argue that a method similar to the method of proofs and refutations is present in scientific as well as mathematical conceptual evolution. Even our best scientific concepts are open textured and ill-defined in certain domains, and science can only grow and change by use of a method that exposes those areas of ill-definition when they occur within models: thought experiments. Second, I develop an account of what makes a monster a monster. Lakatos and Waismann treat monstrosity and ambiguity as a matter of terms, their definitions, and their logical consequences. My account will instead characterize monstrosity as a failure of operational coherence between one or more models. Since scientific models are mathematical or linguistic structures that instantiate scientific concepts, their texture is doubly open. Monsters lurk behind every corner. And sometimes, those monsters can only be found when models are mashed together in new ways and in new contexts. Combined, these two pieces form an account of the progress of scientific concepts – of how science does its

[18] In a set of comments on my paper 'Mathematical SETIbacks' at the Canadian Philosophical Association in 2023, Stephen Ross argued that even chess has exhibited some open texture in its history. I'm grateful for the comment, and I look forward to future research in which I will assess the truth of this claim.

[19] As I have argued elsewhere, whether alchemical gold is considered gold or not was a matter of profound importance within 17th Century natural philosophy. The question of the genuineness of alchemical gold intersects with many pressing metaphysical and theological questions of the period, such as the existence of substantial forms, the relationship of primary and secondary qualities, and the status of teleology. See (Whyte 2021) for a study of one such case.

own conceptual engineering. Though I do not claim that the method I lay out here is the *only* method by which science checks its models and modifies their constituent concepts, I hope to show with some intriguing examples that the thought experimental method has a long and proud history as a tool for improving scientific concepts.

1.2.2 On the Operating Table

The distinctive 'experimental' character of thought experiments comes from the fact that the model engineering qualities that thought experiments test are not typically clear outside the context of a model's use. Where the incompatibilities are obvious, thought experiments are not needed. Thought experiments 'test' the representational structure in question by forcing it to account for an unusual or difficult case[20], but one that is relevant to how the representation would hypothetically be used for its true purposes. As previously discussed, representations are not good or bad *simpliciter*. They are good or bad for a particular purpose in a particular context. Therefore, to determine whether or not a representation serves the purpose we have in mind, we must actually attempt to use it for that purpose. It is one thing to guess that a car with four wheels ought to be drivable. It is another to test-drive it.

Models are non-linguistic entities, so it cannot simply be the case that incompatibility is equivalent to contradiction of the form 'A and ~A'. The unifying feature of model incompatibilities is what some Kantians call Practical Contradiction (as in Korsgaard 1985), and what Hasok Chang

[20] This approach is an expansion of an idea first described by Thomas Kuhn in 'A Function for Thought Experiments' (Kuhn 1977). Kuhn's Function, however, is more limited than what I am proposing here. He claims that thought experiments can generate anomalies in theories, which accumulate and lead to crises over time. However, this account is not sufficient to handle for thought experiments that seem to generate positive results, of which there are many (see Chapter 3 for an extended example of one such positive thought experiment). Kuhn only wrote the one paper on this topic. I see my project as taking up and greatly expanding on this idea.

has termed 'Operational Incoherence'. Operational Coherence or Incoherence is a property of a system of practice, like the practice of modeling a given entity in a given way. For Chang, a system of practice is a systematic, goal-directed activity. Operating a scientific instrument, solving a physics problem with a free-body diagram, or driving a stick-shift vehicle are systems of practice. Every system of practice has a goal, and an operationally coherent one allows its practitioners to achieve its goal, *ceteris paribus*. An operationally incoherent system of practice is one that is ill-suited to the purpose for which it has been designed. It undermines its own goals. If my system of practice for walking down the street involves taking two steps back for every two steps I take forward, that system of practice isn't coherent with my goal to get to my lecture on time. Similarly, if my chosen system of practice for getting to my lecture on time means that I have to traverse the five-kilometer distance between my apartment and the lecture hall in three minutes, that too would be incoherent with my goal, since I cannot reasonably hope to walk that fast. The relocation of these operations to the mind, rather than the streets of Pittsburgh, does not change the fact that some systems of practice are better and some are worse at achieving their goals. If my geometrical practice depends strongly on my ability to square the circle, my geometrical practice will fail. If my attempt to model some object of scientific interest involves modeling it as a round square[21], my model will not help me gain any descriptive grip upon it. The goal of a system of modeling practice is typically to facilitate a particular kind of surrogative inference for a particular kind of inferrer. An operationally incoherent model, then, is a model that undermines the surrogative inference in which its user is

[21] That is, a round square in Euclidean geometry. As John Norton has pointed out, even this, the most cliché of impossible things, is possible in non-Euclidean geometric contexts (Norton 2022). And indeed, I am not here suggesting that the round square wouldn't be good to use because it's impossible, but because a round Euclidean square is *unthinkable*.

interested. Incompatibility is thus a practical problem rather than a formal problem[22], even though its source is the formal structure of a representation.

1.2.3 The Bestiary

Monsters are by their nature unruly creatures, and their homes within specific systems of practice and contexts of use means that they resist taxonomy. However, in this section I will provide some examples of common varieties of monster that arise in modeling contexts and can be discovered by thought experiment. This bestiary is non-exhaustive but hopefully indicative of the diversity of the monstrous regiment.

Formal monsters are the most familiar form of monster to the logician: they undermine a practice by delivering two incompatible answers to the same question, like a contradiction does. This is purely structural incompatibility and would persist even if the concepts used to construct the models were changed. If a model, when used, answers a relevant question with 'A and ~A', that model is incompatible with the way it is being used. If the way it is being used is otherwise desirable, then it is evidence that this is not a good model for that purpose. A good example of a thought experiment revealing syntactic incompatibility is the classic case of Galileo's thought experiment concerning falling bodies, which demonstrates that Aristotle's law of fall can give conflicting answers to the same situation[23].

[22] Representations can also exhibit the formal appearance of incompatibility without thereby exhibiting that incompatibility, since the content of incompatibility is the practical, rather than the formal, component. This is one of the reasons that apparent paradoxes do not always destroy the theories in which they appear, and how quantum field theory has got away with using inconsistent mathematical representations for decades without it really being a problem.
[23] See (J. R. Brown 2010) For a classic telling of this tale, and (Gendler 2004) for a compelling alternative.

Conceptual monsters exhibit incoherence at the level of the meaning of the concepts involved in the construction of a model. These monsters can look like contradictions, but they would not remain a contradiction if the relevant concepts were changed, modified, or swapped out. A good example of this kind of monster is Schrödinger's famous thought experiment featuring a cat in a quantum box trap. The whole point of the thought experiment is that a living cat and a dead cat are incompatible in superposition, but a spin-up particle and a spin-down particle are not so incompatible. The thought experiment thus shows the absurdity of extending quantum explanations into the macroscopic world. The living and dead cat are incompatible because that's just not how life and death work and that's definitely not how cats work. The problem isn't the syntactic structure, the problem is the meaning of 'life'. And, I suppose, the meaning of 'cat'. These monsters are the ones that Lakatos encounters in the polyhedron concept, and the ones that Waismann was concerned about.

Some monsters, if taken into our models, would cause them to simply lose track of the primary phenomena we hope that they describe - an incompatibility at the level of reference. This is the vaguest and weakest kind of incompatibility. A good example is Rudolph Clausius's 1851 demonstration of what would become the Second Law of Thermodynamics. Clausius' thought experiment showed that assuming the non-uniqueness of the efficiency of an engine (over a given temperature difference) would allow the existence of a compound engine that could move heat from a colder to a warmer place, which, he claimed, would be in contradiction with "the general deportment of heat" (Clausius 1851). The implication is that whatever kind of thing Clausius' thought experiment picks out, it does not behave in a way that seems relevant to anything we would recognize as the behaviour of heat. So, a model of thermodynamics that places no limits on the efficiency of an engine isn't a model that picks out the phenomenon we are interested in talking about when we discuss heat. This kind of incompatibility is often vague and hard to define, which is

why models that exhibit it are sometimes just called 'unintuitive' without greater specificity. Many thought experiments in philosophy demonstrate incompatibilities of this kind, which is why they are often so hotly contested.

1.2.4 Where the Wild Things Are

John Norton's rough-and-ready but oft-cited preliminary characterization of thought experiments in his 1991 paper on Einstein's thought experiments describes them as "arguments which: (i) posit hypothetical or counterfactual states of affairs, and (ii) invoke particulars irrelevant to the generality of the conclusion" (Norton 1991). Though Norton intended this characterization to be nothing more than a brief indication of the type of cases in which he was interested, and certainly not a definition thereof, it has persisted as a quotable paradigm within the broader literature on thought experiments. Whatever thought experiments are, they involve the invocation of non-actual scenarios, and those scenarios can contain fanciful details that seem irrelevant to the point the thought experiment is making. My characterization of thought experiments similarly involves non-actual scenarios – imagined use of scientific models in various imagined contexts. Yet, it doesn't seem to be the case that just any imaginary scenario would serve the function of a good test, nor that the scenario described in a thought experiment must never have occurred. A non-actual scenario might be sufficient to make a thought experiment a thought experiment, but not thereby a good one. Likewise, thought experiments sometimes invoke scenarios that are non-actual at the time in which they are being considered but are later realized[24]. What characteristics of an imagined scenario make it appropriate for the function I claim thought experiments serve?

[24] One obvious example is Galileo's falling bodies thought experiment's invocation of the at the time impossible idealization of zero air resistance. The first partial vacuums were still decades away, and the test of Galileo's thought experiment in a real vacuum centuries away. For discussion of another example, to which we shall return in Chapter 4, see (Arthur 2012)

As J.R.R. Tolkien argued in his essay 'On Fairy Stories', quoted above, fantasy can be unlimited in its scope. It need not be constrained to the affairs of mortals like us. And yet, in all the great fantasies and legends, the ones that really shape our society, we find beings much like ourselves behaving in ways much like our own. Even Tolkien's own contributions to the genre have more to say about humanlike hobbits than the otherworldly elves they meet. This, Tolkien claims, is no accident. Though we could very well write fantasies about elves alone, we rarely do – and even less often do we find such stories interesting[25]. The ones that matter to us are the ones about the intersection of the human and faerie worlds. It is in that overlap that the distinctive power of fantasy can be found.

Thought experiments are fantasies in the same sense. To analyze the scenarios they describe as possible or impossible is to mistake their purpose. The question of import is whether they belong to our world or to another. Thought experiments set in unaltered reality alone are of limited interest since they merely describe what is already known to us. Thought experiments involving fully fantastical worlds alone are of no interest because they have nothing to do with us. The thought experiments we care most about are the ones that extend our world by contact *with* one of those others[26]. A world in which there is no air resistance is just as fantastical and impossible as anything to do with elves, but there's a reason for why we see scientific thought experiments invoking the former and not the latter.

Very little is gained, I claim, by placing metaphysical limitations on the kinds of scenarios thought experiments may legitimately depict. A simple pragmatic consideration can do the work of a

[25] Silmarillion fans don't @ me

[26] By the human world I mean not any metaphysical notion of the actual world, and by other worlds I do not mean the other possible worlds of which the metaphysicians speak. After all, Tolkien is not in the business of establishing whether or not Middle Earth is logically consistent or what the Best System of its laws of nature would be. The author of this dissertation is studiously neutral on the metaphysical status of other worlds.

thousand metaphysical distinctions. Pointing out that no ship could possibly sail smoothly enough that Galileo's hypothetical cooped-up sailor would fail to detect its motion by any experiment made below decks (Galilei 1962) makes no difference to Galileo's conclusion. The thought experiment demonstrates that using the same laws to describe a system in uniform motion and a system at rest generates no monsters – and that's true regardless of whether such a uniform motion can be found anywhere on the non-uniform earth. The result of the thought experiment is extremely useful despite its impossibility. It allowed Galileo to model motion in a much more perspicuous way, and the impossible idealization was permissible because it was relevant to precisely the question in which Galileo was interested. Galileo wanted to test whether the laws of motion he had developed for rest would work for uniform motion, so the relevant thought-experiment scenario is precisely a system in uniform motion. No context of use is ruled out *a priori* – the scenarios we imagine for our thought experiments just are the scenarios in which we are interested. A context is relevant simply when it is a context in which we might want our models to operate.

1.3 Regimenting Monsters

In the first section of this chapter I defined a problem for scientific modeling: scientific models are made out of scientific concepts, which are constantly shifting as models are made and used, but the usefulness of models depends sensitively on the circumstances of their use and the other models with which they must always be used. Taken together, these three points imply that any confidence we have that our models actually work well together ought to evaporate every time we try to extend or change any one of them. The vast edifice of scientific modeling is built on shifting sands, threatening to fall into paradox and incoherence at any moment. This problem cannot be solved empirically - laboratories provide the parameters by which the success of answers

can be judged, but no laboratory can expunge the contradictions from a species concept or solve a paradox.

And yet, we do not see science in as dire a state as that which I just described. There must be some way by which science manages to iron out the wrinkles in its patchwork, some checksum that ensures a new application of a model functions as intended, some method by which scientists may reassure themselves that their empirical efforts will not be in vain. We must have some way to check the engineering of our models, not just the application. Some process must serve this function.

I propose that the process that serves this function is the *thought experiment*: a process by which a reasoner imagines, calculates, or simulates the application of the models in question to the context in which they are interested. This process, if properly executed, produces one of two results: either the reasoner finds that they are able to think through the application of the model to the context without trouble – the operations are smooth and coherent, no equations explode, no contradictions are generated; or the reasoner discovers a monster – a paradox, a value that goes to infinity, two answers to a question that ought only have had one, or simply a place where a concept will need to be stretched lest it break. Any non-empirical process that serves to check the coherence of models in this way is a thought experiment on this account.

I take this model-checking function to be definitive of thought experiments. This definition is not found within the existing literature on thought experiments – it is a redefinition of the term that I am imposing upon that literature. The literature on thought experiments rarely gives a firm definition of what thought experiment is[27]. Its authors typically prefer to start from a set of canonical examples (of which Galileo's falling bodies thought experiment is by far the most common) and

[27] Even Norton's brief characterization (given above) is merely designed to provide what Norton considers to be necessary conditions for a thought experiment, with no pretense to sufficiency. (Norton 1991)

account for them alone. I believe that most if not all of the canonical examples of thought experiments within science and philosophy are well-characterized by the account I have just given. However, my definition also includes a much broader set of cases than one would find in a typical list. If I'm correct, thought experimenting is a near-constant part of the activity of modeling. The canonical cases upon which the literature focuses are merely the most charismatic examples of this rather mundane mental process. Galileo's and Einstein's thought experiments are as famous as they are because the results they produced were sufficiently surprising to be written down and widely read. They are the charismatic megafauna of the wild world of thought experiments. The process that produced them, though, is no more arcane than the bored physics student's pen and paper test of what would happen to Newton's laws if gravitational force was an inverse cube law rather than an inverse square. That the same definition of 'animal' that we use to name the elephant also applies to the tardigrade is no mark against the definition.

My account provides a bridge between philosophical accounts of scientific models and the pre-existing literature on thought experiments by giving thought experimentation a specific job within an account of models. In doing so, I have also implicitly limited the strength of a thought experiment's conclusion. Nowhere in the foregoing account do I describe a thought experiment as revealing that some putative fact is true of the actual world – only that a given model can or cannot resolve a given hypothetical phenomenon. Thought experiments are tests of model engineering, but not of model accuracy, and a well-engineered model can still completely fail to be accurate. This account of thought experiments is thus non-evidential – a good thought experiment need not generate any true claim, and a bad one is not bad because its conclusions are false. This too is a significant departure from the extant literature on thought experiments, which is largely divided into two camps: those who think thought experiments are evidential and do generate true claims, and those who think thought experiments have no significant function in science. In the next chapter, I

will put my account to the test against its competitors. I have shown here what a thought experiment can do. Now let's investigate what a thought experiment cannot do.

2.0. Can you Picture That?

"Let me take your picture, add it to the mixture, there it is I got you now
Really nothin' to it, anyone can do it, it's easy and we all know how
Now begins the changin', mental rearrangin', nothing's really where it's at
Now the Eiffel Tower's holdin' up a flower
I gave it to a Texas cat
Fact is there's nothin' out there you can't do
Yeah, even Santa Claus believes in you
Beat down the walls, begin, believe, behold, begat
Be a better drummer, be an up and comer
Can you picture that?

Can you picture that?"

- "Can you Picture That?", Dr. Teeth and the Electric Mayhem

In this chapter I will argue that thought experiments are a method by which science changes its models, and therefore, its concepts. In describing thought experiments as a model engineering method rather than as either a source of new knowledge about the world or an elaborate way of presenting arguments about the world, I deviate from most of the extant thought experiment literature. I will contend that my account of the function of thought experiments not only solves the problem identified in Part 1, but a number of outstanding problems in the literature on thought experiments as well. I do not claim that the method of thought experimenting is the only way that science generates and improves its models. However, I hope to demonstrate that this strategy has a long history of success. Thought experiments generate the natural scientific analogue to the 'monsters' in Lakatos' account of mathematics – domains of application of our models that reveal their open texture and force us to make a choice about how they ought to be properly used.

After that, I will argue against other accounts of thought experiments in the literature. I am going to argue that the root problem of all these accounts of thought experiments is the same: the assumption of evidentiality. The assumption of evidentiality is the assumption that a successful thought experiment is successful *when and only when* it generates evidence for a true fact about the world. The philosophers who make this assumption take it on as a defense against critics who regard thought experiments as idle fantasies, useless pretentions, or merely rhetorical exercises (cf. (Dennett 2013), (Thagard 2014), and (Machery 2011) for notable contemporary versions of this critique). That is not my criticism. The aim of my account is not to deny science access to the fruits of the imagination. I do think that thought experiments have a pivotal role to play in science, and I think any adequate history of science will clearly show that role. Indeed, in subsequent chapters I will illustrate cases of thought experiments in their historical contexts and demonstrate how vital the thought experiments were within those contexts. However, science is not merely composed of fact-gathering, and activities that do not themselves generate novel facts can still be crucial to the enterprise of science. In this section, I will argue that the usefulness of thought experiments does not depend upon their fact-generating powers, and that indeed, they have no such powers.

In short, there are two main arguments I hope to make in this section: a positive argument about what thought experiments can do, and a negative argument about what they can't do. Then, I will work out some consequence of this account.

2.1 What is required of an account of Thought Experiments?

In the previous chapter, I described a role for a vital process within science, and I named that process 'thought experiment'. I did not choose that name at random, and I am not the first to use it. The term has a long history of its own, and in this section I will argue that my use is adequate to that history.

The term 'gedankanexperiment' is believed to have been coined by Hans Christian Ørsted in 1811, but it owes its popularity to the writings of Ernst Mach and Albert Einstein, and its honour to Einstein's many famous uses of the term in his revolutionary early work[28]. Thought Experiment as a genre, then, was born in triumph. Einstein changed the world, and he did it with a mere imagining. The concept 'thought experiments' was then read back into other famous imaginary cases in the history of science and philosophy, from Galileo's three interlocutors and their days discussing hypotheticals, to the myth of the Ring of Gyges that Plato relates in the Republic[29]. And once the notion 'thought experiment' became commonplace amongst philosophers, it became a common philosophical self-description of philosophical activity. Google's Ngram analysis of English-language writing shows the term 'thought experiment' was very seldom used until the late 1950s. It grew slowly in use throughout the 60's and 70's before exploding in popularity in the 1980s. Thus, there are three distinct ways that the term 'thought experiment' gets used in even the canonical set of thought experiments discussed by the literature: self-description of scientific practice with no honourific subtext, retroactive description of historical cases, and self-description of philosophical and scientific practice with honourific subtext. The canonical set of thought experiments is drawn from a mix of much-lauded historical exemplars and modern mimics of those exemplars.

The result of this mixed history is that the term 'thought experiment' refers typically to a wildly heterogeneous set of cases. The canonical thought experiments belong to many disciplines, many time periods, and many publication contexts, with the term 'thought experiment' itself acting as both an actor's category and an analyst's. Though the prototypical historical thought experiments are remembered because they were highly successful, the success of later examples that carry the

[28] Though, as Sara Roux points out, this story is somewhat hackneyed and simplified. See (Roux 2011) for a more detailed and critical account of the emergence of the term.
[29] Some of this assimilation of the notion of thought experiment to Galileo was done by Einstein himself, as Roux (ibid) notes.

thought experiment banner is hotly contested. 'Thought Experiment' is used to distinguish a kind of practice more epistemically noble than mere fancy, but it is not a success term. I put forth that the messiness of the phenomenon we call 'thought experiment' leads to a messiness in purported accounts thereof. Most authors who discuss thought experiments have been content to take an "I know it when I see it" approach to their subject matter, and I think this lassitude causes more problems than it solves. It is hard to make either a normative or a descriptive account of a phenomenon when you don't know what you're describing or what norms it should obey. So, let me take a metaphilosophical moment to lay out terms upon which I believe a debate about thought experiments ought to be conducted.

2.1.1 Desiderata for an account of thought experiments:

An account of thought experiments must provide a positive view of a) what thought experiments do and b) how they do it. The account should provide criteria of success and failure for thought experiments, and also clear criteria for their identification (or at least, their identification in context of sufficient historical evidence). It is also desirable for an account to be able to give some explanation for why thought experiments occur when they do in history.

These criteria are not typical of the literature, so I should defend them a little. At this point in the literature on thought experiment there is not much more to be gained by simply providing examples of things that could be called thought experiments, or by multiplying senses of the term 'thought experiment'. Unless a researcher is working with an account that can wield some actual descriptive or epistemic force, arguing that 'Case X is an example of a thought experiment' says little more than that it shares some feature or other with the now very large set of canonical cases. Likewise, the term is used currently to refer to such a wide variety of phenomena in such a broad set of ways that merely pointing to a new way in which something can resemble previous cases again

says very little. It is also no longer possible to do an analysis of thought experiments by analyzing the use of the term 'thought experiment', since it has passed into general use in the English language. The example base is far too heterogeneous, both temporally and historiographically, to allow any analysis of the use of 'thought experiment' as a phrase to be philosophically enlightening. A theory of thought experiment must, therefore, be a theory of some phenomenon picked out by the term in a typical case, and as a consequence, must allow both that it is possible to use the term incorrectly, and that it is possible the term has been used incorrectly elsewhere in the literature. It cannot merely be a record of the many ways in which 'thought experiment' is said. The point is not to use our linguistic intuitions, but our philosophical judgement.

Moreover, it is vital, I think, to be able to make sense of both the success and failure of thought experiments. It is easy to focus on success in the history of science. Success gets better press. The so-called 'file drawer problem', in which scientists choose not to publish results that didn't work out as expected, likely predates the existence of file drawers and the existence of anything going by the name 'science'. The most celebrated examples of thought experiments are also celebratory examples, and that is what the bulk of the literature is about[30]. This bias in the literature is understandable, but it can obscure the subject matter under discussion. 'Thought Experiment' is not a success term. If thought experiments perform any kind of function in science other than the purely decorative, they must be capable of failure. We should be able to account for those failures in our analyses of thought experiments, and we should be able to say why and when they happen.

I believe that these desiderata – that an analysis of thought experiment cannot be an analysis of the use of the term 'thought experiment' and that an account of thought experiment ought to show that and how thought experiments both succeed and fail – push towards an account of

[30] There are exceptions. See (Norton 2018a) for one.

thought experiment in which the term is defined in terms of a function. Attempts to satisfy that function can either succeed or fail, and so long as sufficient historical evidence exists, they can be unambiguously identified. So, what function?

2.2 Can you picture that?

In the previous chapter I established three central claims: Sciences changes over time, the basic unit of scientific change is the scientific model, and a scientific model is only good iff it can be productively used to represent some narrow sliver of the world, which it must do in the context of other models. These three conclusions lead us to a problem over time. If scientific models are constantly shifting partial descriptions of a dimly glimpsed world, then how do we ever know if our current ephemeral combinations of models actually work together, and thus whether they can be used for the purposes to which we put them?

My answer is that we can learn when our models can be used together when we can successfully 'think through' their use in a situation relevant to the ones in which we'd like to be able to use them. We put forth a hypothetical case and then try to work with it as if it were a real case. We operate with the affordances the model provides unless we hit a paradox or snag. If we can coherently apply the different aspects of our models without generating paradoxes or nonsense, we know the model is well-enough engineered to be used in the context in which we'd like to use it. If the models work harmoniously together to deliver an answer to the hypothetical case, they are compatible. Sometimes thought experiments test representations against themselves, sometimes they test them against others. Thought experiments are thus procedures, not entities.

Though thought experiments render judgment on questions of the compatibility or incompatibility of our models with themselves and each other – Maxwell's Demon demonstrates the *compatibility* of the statistical approach to thermodynamics (which allows for entropy to decrease, but improbably) with phenomenal thermodynamics (which never allows entropy decrease), Schrodinger's cat demonstrates the *incompatibility* of a unified description of classical and quantum states – I argued that they are not sources of knowledge about the world. Conceivability in a thought experiment implies *applicability*, but it does not imply either possibility or actuality[31]. Neither of these cases give us insight into the true or possible nature of the world, only the adequacy or inadequacy of our methods of representing it in certain cases. Maxwell's demon describes an impossible supernatural scenario, and the true upshot of Schrodinger's cat for the world remains unclear almost a century later. But both succeed in showing the power and limits of their respective modeling strategies: the statistical formulation of entropy on the one hand and the 'cut' between quantum and classical on the other.

I think that this model-testing function is the defining quality of thought experiments, and that any mental process that is undertaken in order to perform this function should be considered a thought experiment. Defining thought experiments by function rather than by the distinctive phenomenology of canonical examples brings thought experiments more in line with laboratory experiments, which are also identified by epistemic function rather than the mere presence of Bunsen burners and Erlenmeyer flasks[32]. It also satisfies the desiderata outlined above. A functional definition allows us to unshackle ourselves from mere use of a term, it provides conditions of

[31] Some accounts of thought experiments, most notably Roy Sorensen's account, claim that thought experiments primarily reveal whether certain theoretical claims are true, physically possible, or metaphysically possible. I shall not dwell long on Sorensen's account, but I think it is sufficient to say that an account on which it is possible that every thought experiment in history has been unsuccessful is not a satisfying account of the phenomenon of thought experiments given the criteria above. See (Sorensen 1992) for the positive account of this view.

[32] For instance, the Stanford Encyclopaedia of Philosophy article on 'Experiment in Physics' opens by simply listing the functions that experiments can have within physics (Franklin and Perovic 2023)

success and failure (the satisfaction or lack of satisfaction of the function) and it allows us to clearly identify thought experiments within history so long as we have adequate evidence for the intended function of the relevant cases.

Let me present the account more succinctly and clearly:

2.2.1 The Model Engineering Account of Thought Experiments

A Thought Experiment is an **abstract procedure** by which a **thinker** puts their **methods of representing** some **aspect of the world** to a **test** in order to determine whether it is **feasible** to use that method to represent that aspect of the world in the novel **circumstances** of the test. Every thought experiment has one of three outcomes. If the thought experiment is **well-formed,** the experimenter will discover that the methods of representation are or aren't **coherent** with **use** in the circumstances of the test. If the thought experiment is badly formed, the result will be **inconclusive**.

2.2.2 Exorcising the Details

In this section I will walk through all the bolded terms in the account given above and clarify their meaning in the context of this account.

'**Abstract'**. Thought experiments are experiments on the contents and structures of thought, not on physical objects. This is what makes the difference between thought and laboratory experiments. However, despite the infelicitous nomenclature, a thought experiment need not be carried out only (or even primarily) in the mind. Many canonical thought experiments involve pen-and-paper methods or illustrations. The function I posit for thought experiments can also be fulfilled by some

(but not all) uses of computer simulation. All these non-mental forms of abstraction are allowed by this account.

'**Procedure'**. A Thought Experiment is not an entity. The words that convey a canonical thought experiment are not the thought experiment itself, and neither is the fictional scenario that is worked through in the thought experiment. Identifying the thought experiment with its fictional scenario is a mistake for the same reason that it would be an error to identify a laboratory experiment with one of its Bunsen burners. Likewise for the textual presentation of the thought experiment and a laboratory report. A thought experiment is the same kind of thing as a laboratory experiment – a procedure.

'**Thinker'**. There is no domain-specificity in my account of thought experiment. The same account applies to philosophers and scientists engaged in the thought experiments of their domains. This lack of division may strike some as distasteful. It is common for philosophers to bemoan the poor state of evidence gathered from the thought experiments of their own discipline but allow that thought experiments could be legitimate sources of evidence in the securer sciences[33]. These philosophers are correct that thought experiments provide a poor evidential basis upon which to ground their theories, but they are wrong, I claim, in suggesting that the science should fare any better.

'**Methods of Representing'**. Virtually every activity that science performs save for experimentation itself is done using representations of the phenomenon in question rather than the phenomenon

[33] This sharp division in both power and respectability of scientific and philosophical thought experiments can be attributed to Kathleen Wilkes' influential book on the subject of thought experiments in personal identity. I do agree with her contention that these thought experiments are often critically ill-defined in comparison to their scientific counterparts, however. See (Wilkes 1988) for this critique.

itself. Surrogative reasoning is the core of scientific representation. Thought experimenting, in my account, concerns only these surrogates[34].

'**Aspect of the World'**. Representations have to be representations of something, and the something in which we are interested is typically the world in one of its innumerable aspects. Even investigations of possible but non-actual circumstances (eg. a Gödel space-time (Gödel 1949)) are investigations of aspects of the world, since they are pertinent to our understanding of the world in which we actually live. This distinguishes a thought experiment from pure fiction but allows that the situations conjured in thought experiments are still fictional. Stanislaw Lem's fable of the Demon of the Second Kind (Lem 2002) is undeniably fictional, but insofar as it is a commentary on how we think about the limits of the concept 'information', it can still give rise to a thought experiment. As Tolkien notes in the paragraph exerpted as an epigraph to this dissertation, we learn nothing from pure fairy stories – we learn from stories about the adventures of men in the perilous fairy realm (Tolkien 1966). Some connection to the world in which we live is necessary for a thought experiment to achieve any task about which we would reasonably care.

'**Test'**. The safety of cars is established by crash-testing them. The stability of materials is established by stress-testing them. The readiness of undergraduates for an advanced differential equations class is established by calculus-testing them. In each case, the subject of the test is subjected to an extreme version of the same kind of situations they would be expected to encounter in the capacity for which they are being tested. A test is good insofar as it accurately sorts the candidates that are good for a particular role from the ones that are not good, without too many false positive or false negative outcomes.

[34] As discussed in the preceding chapter, this account assumes that models are semi-autonomous abstract mediators à la Morgan and Morrison's account of models as mediators (Morgan and Morrison 1999) and a broadly Gierian perspectival account of scientific representation (Giere 2010).

'Feasible'. 'Feasible' here is meant in the sense that it is feasible to divide a number by 2 but not by 0 – not a modal notion of possibility but a reflection of the affordances offered by the method of representation. All methods of representation, be they mental models, systems of equations, physical models, or digital simulations, permit some operations and not others. Sometimes when we try to model some phenomenon using our rolodex of familiar modeling methods, we discover that there is no operation within our modeling method that can render the phenomenon in question, or perhaps that our methods can render the phenomenon in two incompatible ways, or that modeling the phenomenon causes one of our other analytic tools to fail. All of these are sufficient to render the representation unfeasible. Model engineering, as discussed in the previous chapter, is the process of building feasible models.

'Use'. Minimally, the point of a scientific representation is to allow the representer to get some sort of useful handle on the world. These uses are as varied as the systems they represent. In the broadest sense of the word, a representation is good *qua* representation iff it can be successfully used for the purpose for which it was intended. Most of the time, a representation in science is used to predict or control; in philosophy to explain or unify, but purposes vary widely in both disciplines.

'Circumstances'. Most scientific representations are designed with some particular use-case in mind. Some methods of representation successfully undergo domain extension, and some of them do not. Representations, in this way, are like words. They have a natural area of description – the one that they were built to describe. But many of the great victories of the history of science have been those in which new areas have been brought under the same modeling umbrella. All successful modeling strategies have some domain over which they are successful, and the very best modeling strategies (like, say, Lagrangian Mechanics) can subsume entire disciplines under one modeling frame. However, no domain extension is guaranteed to be successful. All ways of representing the

world have hidden points of failure, which cannot necessarily be detected *a priori* because they only emerge in particular circumstances of use. Thought experiments can find these hidden blind spots in our modeling frameworks[35].

'Well-Formed'. A close analogue to the experimental notion of construct validity. A well-formed thought experiment is free of vagueness (beyond the vagueness it might expose) and concerns the relevant aspects of the circumstances investigated. In short, a thought experiment is well-formed if it truthfully reveals whether or not the method of representation in question can be used in the circumstances of the test. This is the success condition of a thought experiment.

'Coherent'. What is the standard for whether or not a method of representation can represent a given target? A critic might argue that any method can be used to represent any target, just with greater or lesser efficiency. This is only half-true – representing a discontinuous Heaviside step function as a limit of smooth logistic functions, for instance, can be useful in some applications but cannot replace the discontinuity in others. More importantly, most methods of representing cannot represent most targets well or usefully. We have good reason to reject methods of representing that are inefficient, cumbersome, inexact, and difficult to use for the purpose we have in mind. Chiaroscuro illustration is a very ineffective way of representing spacetime, and Penrose diagrams are a very effective way of doing the same. Game theory is an excellent way to understand the behaviour of professional gamblers, less so the behaviour of toddlers. Judging a representation to be incoherent means judging it to be either incapable of or ineffective at representing the target in a fruitful way.

[35] This point is drawn from (Lakatos, Worrall, and Zahar 1976) and (Polya 1954a). Their focus was on the extreme case of mathematics, but I believe the extension into the physical sciences (and then into concepts in general) is very natural, however ironic it may be.

'Inconclusive'. Any experiment can fail, and thought experiments are no exception. If the thought experiment was ill-formed, it will render no result at all. This null result is analogous to Quine's sense of 'falsidical paradox' – a conceptual mistake or misapplication (Quine 1966). Thought experiments that trade on equivocation are the most obvious form of ill-formed thought experiment, but there may well be infinitely many different ways to be useless.

2.2.3 Consequences of the Model Engineering Account

Nothing in the foregoing account of thought experiments allows that thought experiments can judge whether any claim about the world is true or false, nor possible or impossible. This is because thought experiments are fundamentally not about the contents of the world at all. They are about the tools we use to describe that world. To return to the distinction between model engineering and model application invoked in the previous chapter, thought experiments are tools of model engineering and silent on model application.

The literature on thought experiments has always been partially hamstrung by the apparent continuity between the distinctive activity of thought experimenting and other uses of fiction, models, and narrative in science and philosophy. Humans are creatures of magnificent imagination, and that imagination has a tendency to creep into nearly everything we do. This is a problem for philosophers like the proponents of the theories I will discuss below, who desire to hive off thought experiments as a method of inference from the other uses of the imaginative faculty in science. In each case, the distinction seems artificial – the question of what 'really is' a thought experiment establishing a barrier where none is present in thought. If our goal is to understand current and historical scientific reasoning, the artificiality of this distinction should worry us, since it forces the

philosopher to adopt a divide between categories that its practitioners would reject[36]. However, not making that distinction is equally untenable. The use of the non-real in our investigation of the real is so thoroughly ingrained that it is hard to recognize in all its guises. The most basic forms of scientific representation, like the representation of weight by a positive scalar quantity or a beam of light by a line, are so natural to our way of scientifically approaching the world that we scarcely even notice that they are fictions. Light is not a line, it is light; weight is not a number, it is weight. Should we think of all classical dynamics and all geometric optics as thought experimental? Only if our notion of thought experiment is so weak as to be contentless.

The intuition I am trying to evoke by these examples is that identifying thought experiments with the legitimate presence of fantastical elements in science leads to a loss of the distinctive phenomenology of the thought experiment, which lies not in the thought, but in the experiment. If, as I have proposed, there is a necessary experimental character to thought experiments, then we need not identify being a thought experiment with merely the content of the relevant thoughts. This would be analogous to identifying a laboratory experiment with its materials. This is why I have proposed that 'thought experiment' should be a name not for a kind of imagined scenario, but for a kind of activity. The theories I will discuss in the critical section below parse 'thought experiment' as an experiment *within* thought. In the account I promulgated above, 'thought experiment' refers to an experiment *on* thought. Thought experiments are not experiments performed upon the world using the apparatus of thought, but experiments performed upon our methods of thought. The results of our thought experiments are answers to questions about our thoughts themselves.

The explanatory burden I must shoulder in arguing for this account is that the canonical stable of thought experiments certainly seem to have been very useful at major turning points in the

[36] And, moreover, about which historical evidence is necessarily thin on the ground. See (Stuart forthcoming) for a summary of these difficulties.

history of science. They are strongly historically correlated with new theories, and true ones too. If, as I have argued, their conclusions can only be about the thoughts of the experimenter, then I must explain how conclusions about such airy nothings can be as active a force in science as they undeniably are. Thought experiments have a perennial place in the history of science and philosophy, from Aristotle to Einstein. Integrated History and Philosophy of Science need not commit itself to the *goodness* of all the multifarious methods that history shows us, but it does require that we understand why certain methods have the staying power that they have. Few have the staying power of thought experiments. Indeed, even the presence of the word 'experiment' disguises the antiquity of this method, since thought experiments are older than the more paradigmatic laboratory experiments. Thought experiments are found in the oldest and newest scientific writing – sometimes even the same thought experiments feature in both[37]. Scientific methods in general have changed a great deal since the time of Aristotle, but thought experiments seem to be largely the same. This is because, as I have argued, thought experiments serve a function to science that science can never outgrow.

I believe that the account I have given accords with the desiderata I gave before. The Model Engineering account of thought experiments provides an account of what thought experiments do (test models for coherence) and how they do it (simulated use). It provides criteria for the success of thought experiments (accurately revealing whether a model can render the phenomenon at issue) and allows that they can still fail. My account even provides some implication of where we should expect to see the most thought experiments – context in which scientific representations are

[37] We shall investigate Aristotle's Wheel, an example of a single TE that has persisted in scientific and philosophical literature from its probable inception in the 3rd century BCE until the present day, in a later section.

changing[38]. Of course, I am hardly the first to propose an account of thought experiments. Let us now turn to some of the other contenders in order to see how they fare.

2.3 Consider the Ogopogo: Puzzles for Thought Experiments in Science

My account of thought experiments is crucially non-evidential. Most positive accounts of thought experiments in science are evidential. In this section I will argue that any account of thought experiments that claims that thought experiment generate facts about the world needs to solve a problem that may be insoluble. So, for the moment, let us take on the assumption that thought experiments do reveal true facts about the world, and see how far that can take us.

In the interior of British Columbia there is a lake called Lake Okanagan, and in Lake Okanagan lives the Ogopogo, or so the local legends say. It is easy to picture the creature lurking in the lake, with its vast, serpentine body, green scales, horse-like head, and golden eyes. But the Ogopogo is not real. None of my imaginings about the Ogopogo are true, and my ability to imagine them has nothing to do with their truth or even their possibility.

Once upon a time in a world without air resistance, Galileo dropped two balls, a heavy cannonball and a light musketball, off a tall tower. Aristotle's law of fall claimed that the heavy ball would naturally fall faster than the light ball. But when Galileo imagined tying the two balls together he found something strange: Aristotle's law of fall could not tell him what would happen. Would the combined weight of the two balls make both fall faster than either individually, or would the light ball's slower natural motion act as a parachute to the heavier one, causing both to fall more slowly than the heavy one would on its own? The only way out of the puzzle, the sage voice of Salviati

[38] My historical investigations bear out this prediction, but evaluating the distribution of thought experiments throughout the entire history of science and philosophy is beyond the scope of this project.

claims, was that the balls must naturally fall at the same rate, so neither could speed up nor slow down the other (Galilei 1914). The independence of mass from speed of descent would serve as the foundation for Galileo's celebrated law of fall. None of the entities imagined in Galileo's story were real. Even if musket-balls and cannonballs and towers are real, the idealized ones that Galileo invited us to imagine were no realer than the Ogopogo. Yet, when Galileo imagined his scenario, he somehow ended up on the other end with a remarkable scientific breakthrough – that Aristotle's law of fall could not be the case[39]. What makes this case different from the one where we imagined a lake monster?

The classic problem of thought experiments is the question of how merely imagining a scenario, especially a scenario that is impossible, could provide us any true knowledge about the external world. After all, it is easy to imagine things in ways that do not provide true knowledge of the external world, as I did when I conjured the Ogopogo above. Further, when we do find the truth in one of these fantastical journeys, where does this truth come from? No new empirical evidence was granted to Galileo in his imagining, just as my vision of the Ogopogo does not constitute a sighting. Moreover, it seems as if no empirical evidence is needed to further confirm Galileo's conclusion, nor any possible empirical evidence disconfirm it. It is possible to repeat in reality the experiment that Galileo conducted in his mind in a real vacuum, as the Apollo 15 astronauts did on the moon, but it would seem strange to argue that Galileo's result was uncertain for the intervening 400 years. This puzzle has led philosophers of science to speculate about whether or not the intuitive grasp of phenomena in a thought experiment represents something different from the usual standards of empirical confirmation that we would normally use in a scientific context[40].

[39] For discussion of this classic case, see Gendler (2000) or Palmieri (2017)
[40] For the positive and negative positions on this point, see James Robert Brown's "Why Thought Experiments Transcend Empiricism" (2004) and John Norton's "Why Thought Experiment do not Transcend Empiricism" (2004). We will discuss both papers later.

2.3.1 Thomas Kuhn and the Ogopogo

It is possible to imagine the Ogopogo. Imagining the Ogopogo does not justify belief in any true claim about nature. It is possible to imagine Galileo dropping balls off a tower and doing so seems to justify belief in a true claim about nature. How does merely imagining Galileo justify a belief? How does imagining the Ogopogo fail to do so? And how can we know which of our imaginary journeys justify beliefs about the real world and which don't?

This puzzle is not entirely novel, of course. It is a close cousin of the problem that has animated the debate around thought experiments since the first discussions of thought experiments as such. Kuhn's statement of the problem is classic:

> "If we have to do with a real thought experiment, the empirical data upon which it rests must have been both well-known and generally accepted before the experiment was even conceived. How then, relying exclusively upon familiar data, can a thought experiment lead to new knowledge or to new understanding of nature?" (Kuhn 1977) page 247.

Kuhn frames the problem as a problem for the *novelty* of the products of thought experiments. Imagination produces no new sensations, no new *ἐμπειρία*, no new contact with the world. Thus, any product of a thought experiment must be composed of recycled materials – it cannot reveal anything that was not already implicit in the concepts we used to frame the problem.

I concur both with Kuhn's diagnosis of the problem and with many aspects of his solution to it. But I think that the comparison between Galileo and the Ogopogo brings out a salient part of the issue not contained in Kuhn's diagnosis: if the reliability of the imagination in general is doubtful, then we should doubt even whether it is even able to re-arrange the familiar data of sense experience into felicitous new shapes. The description of the Ogopogo I gave earlier (green scales, horse-like head, golden eyes) is perfectly picturable to a human imagination. Our ability to picture it derives from our past experiences of creatures with these characteristics (geckos, horses, and cats, perhaps), though presumably not one creature with all of them. Imagining the Ogopogo nets us scarce epistemic benefits even if all the pieces of our imaginary picture were legitimately obtained. So, there are two questions: Kuhn's question of how we can possibly learn something new by mere imagination, and the further question of how and how far we can trust anything learned in this manner, new or not. I put forth that a satisfactory account of thought experiments ought to provide an answer to both questions.

There are a few possible paths towards an answer to Kuhn's Problem and the Ogopogo Problem. Each requires us to indicate a relevant difference between the case in which we imagine the Ogopogo and the case in which we imagine Galileo such that the latter is a source of genuine knowledge and the former isn't. There are three broad ways to solve this problem: First, the way we imagine Galileo's thought experiment could be different from the way we imagine the Ogopogo such that one is reliable and one isn't; Second, Galileo's thought experiment could have an extra logical structure in addition to the imagining that the Ogopogo story lacks that makes up the supplementary justification; and third, there is no relevant difference between the two stories and thus, either both provide knowledge of the world or neither do. The account I gave above supports the third option. Most of the extant literature on thought experiments argues for one (or a combination) of the first two.

My Model Engineering Account solves Kuhn's Problem and the Ogopogo Problem by cheerfully denying that they are problems at all. Imagining the Ogopogo does not fail to teach us about the world in a way that Galileo's thought experiment succeeds – neither tell us a single thing about the world. Both teach us only about the content of our own skulls, abstract and fallible as any dream. The difference between them is a difference in the use to which we put the conceptual results, not a difference in the status of lake monsters or abstract towers. Galileo's thought experiment demonstrates the incoherence of the Aristotelian account of fall that was commonly used in his own time, and establishes the coherence of his own account, counterintuitive as it may be. He goes on to use this new account very successfully throughout the Two New Sciences. Perhaps somebody someday will find a conceptual structure to test by imagining the Ogopogo, but I am not yet aware of one. But there is no more in that distinction than the practical fact that our core conceptual structures have more to say about falling bodies than they do about lake monsters.

Most extant accounts of thought experiments, however, attempt to establish some kind of non-pragmatic bulwark between the Ogopogo and Galileo in order to secure the fact of the latter against the fiction of the former. In this section, I will investigate three of the most prominent accounts of thought experiments, all of which attempt to make this distinction: James Robert Brown's Platonic Account, John Norton's Argument View, and Nancy Nersessian and Nenad Miščević's respective Modeling accounts. I will argue that none of them meets the challenge that the Ogopogo example presents. The three accounts I discuss all fail for the same reason – they claim that the function of thought experiments is to infer propositions about the world, and that a successful thought experiment justifies such propositions. I claim that the function of thought experiments is not inference, and therefore I am not subject to the same concern. Yet, as I have argued in the previous chapter, the inability of thought experiments to justify propositions about the world does not thereby imply that they have no place in science. Thought Experiments instead have

the function of demonstrating compatibility or incompatibility between mental representations of scientific phenomena. This function is not undermined by the possibility of imagining the Ogopogo. Let us investigate each of these accounts in turn.

2.3.2 Seeing the Ogopogo

James Robert Brown's Platonic account of thought experiments solves the problem of the justification of their conclusions in the most direct way possible. Thought experiments produce conclusions that are justified in the same way that ordinary empirical judgements are justified: through perception. Brown claims that a thought experiment gives the mind's metaphorical eye an opportunity to exercise its ability to perceive the universal laws of nature directly, just like its literal counterpart. The peculiar imaginary scenarios of thought experiments are windows through which the human intuitive faculty can directly perceive the laws of nature and learn *a priori* synthetic truths from them (J. R. Brown 2010).

Most commentators who have argued against Brown's view base their objection to it on the metaphysical and epistemic implications of direct perception of the universal laws of nature. Brown's account would commit a believer to a broadly Armstrongian realist account of laws of nature, and a powerful theory of Platonist perception that allows direct access to these laws. These worries are not sufficient for an argument against this view – most epistemic claims require certain metaphysical commitments, and systematicity is a virtue of a philosophical picture rather than a vice. Indeed, if it were true that we could clearly get *a priori* knowledge of the furniture of the world through this kind of direct perception, that fact would make a compelling case for the metaphysical structures that Brown's view requires. However, I claim that Brown's view does not offer a sufficient epistemic account on its own merits.

A pressing issue with Brown's view is that it provides little insight into how a thought experiment conducted in this way could ever fail. We cannot merely posit that the mind's eye is like the skull's eye without recognizing an analogy between the limitations of the latter with the limitations of the former. The laws of nature are universals, and visions of the forms should admit none of the impediments of crude matter. Brown admits that even amongst the class of thought experiments that provide genuine insight by direct perception (so-called 'Platonic' thought experiments) there are examples that fail to justify their conclusions, like Einstein, Podolsky, and Rosen's infamous attempt to disprove the Copenhagen interpretation of quantum mechanics in favour of a hidden-variables model[41]. A brief overview of the history of science will provide many more examples – both examples of cases in which a thought experiment gave a wrong answer in its own time and cases in which thought experiments seemed to give a correct answer that has since been eroded by scientific change.

As Norton (Norton 2002) has argued, the apparent fact that we can make mistakes when we directly perceive the laws of nature demands an explanation if we are to think of thought experiments as reliable methods of inquiry. Brown's response, that we may justifiably trust our ordinary perception despite both its fallibility and our general lack of understanding of it (J. R. Brown 2010), does not meet this challenge. The worry is not that we lack a full theory of Platonic intuition – it's that we have no knowledge of the kinds of situations under which our Platonic vision is reliable. Even without a theory of vision, it is possible to know that ordinary vision is less trustworthy in dark or foggy conditions, that it can be blocked by opaque objects or blinded by bright ones. In order for thought experiments to be reliable reasoning strategies, Brown must answer the parallel question – how can thought experiments fail to show us the world? What sort of

[41] For a fuller account of the EPR paper as a Platonic thought experiment, see (J. R. Brown 2010).

interference can prevent us from seeing the laws of nature? What does it mean for a thought experiment to fail? In the EPR thought experiment, the thought experimenter purportedly perceives that a hidden variable theory must explain the set-up, given a background of Special Relativity. Yet, that thought experiment's conclusion is misleading. So, in order to explain how these thought experiments can fail, Brown needs to explain why the intuitions invoked by some thought experiment set-ups teach us about the world while others do not. In order to give us normative prescriptions about thought experiments when the truth of the conclusion is not known already to history, we must already know what kinds of imaginary situations can offer us intuitions about real phenomena.

This problem is typically framed in terms of justification – if it is the case that thought experiments provide knowledge in the way Brown describes, the source of that knowledge is at best obscure and at worst unreliable. A source of evidence that isn't reliable and for which the causes of failure are unknown is not a good source of evidence. If we cannot distinguish the good and bad dreams from each other without the benefit of hindsight, we cannot depend on dreams.

2.3.3 Arguing with the Ogopogo

On the other side of the Rationalist/Empiricist divide over the epistemology of thought experiments is John Norton's Argument View. Norton claims that thought experiments render judgments about the world because they are, in fact, fancifully disguised arguments. Norton claims that a careful historian may rationally reconstruct the core argument that sits at the heart of even the most abstruse thought experiment. Those reconstructed arguments are epistemically equivalent to the thought experiment from which they are derived. It's a simple solution to a complex problem - Norton's account deflates the spooky epistemology that was so concerning in Brown's account of

thought experiments to something that seems much more familiar and down-to-earth: the ordinary arguments we use in all aspects of life.

Most critics of the Argument View direct their attention towards one of Norton's two reconstruction premises: that every thought experiment can be reconstructed as an argument, and that the reconstructed argument is epistemically equivalent to its thought experimental source[42]. My approach will be different. I think the reconstruction premises are prima facie plausible enough to grant for the sake of argument, and that Norton's account fails independently of them.

The first premise, that all thought experiments can be reconstructed as arguments, is contingent in its plausibility on the broadness of the meaning of 'argument', but Norton has a very broad notion of argument in mind. He includes not only formalized deductive logic but also an open and contextual notion of inductive logic derivable from his own Material Theory of induction[43] as suitable structures for a reconstructed thought experiment. Reconstructing a thought experiment as an argument seems to always be possible (and Norton has claimed that he has yet to find a counterexample despite decades of searching). The familiar canonical examples of thought experiments are the ones that science decided to write down and repeat, so they always have an expression in ordinary language. Ordinary language is itself reconstructable as a series of more-or-less compelling arguments, as every introductory philosophy student has at some point learned. So, the first reconstruction premise is unproblematic.

The second, that the argument so generated will be epistemically equivalent to the more narrative presentation of the thought experiment in its original source, is more controversial. Opponents of this premise argue that Norton's account necessarily misses the distinctive

[42] See (Gendler 1998) for this critique in full.
[43] Though, as Mike Stuart has argued, Norton's Material theory of induction and his account of thought experiments might be *too* compatible. Stuart shows that the broadness of the material theory reduces the strength of the Reconstruction premises almost to triviality. See (Stuart 2020)

phenomenology of thought experiments, and thus that it misses some amount of the epistemic payoff thereof. Norton's rejoinder is that any epistemic payoff that can't be represented as an argument isn't worth the trouble, since sound arguments are the benchmark for good reasoning. Any account of thought experiments in which they weren't so reconstructable wouldn't be one worth having[44]. If the function of a thought experiment is to prove claims about the universe, as Norton would have, this seems like a reasonable standard to uphold. There's no reason to have an account of bad reasoning! So, I will grant this premise for the sake of argument now.

Norton's strategy with the argument method is broadly deflationary - to reduce the complex mystery of thought experiments to the clarity of argument. Argument is the most basic philosophical tool, and nearly the oldest. Arguments certainly feature in thought experiments and our use of them – I am not arguing that they have no place in an account of thought experiment. However, I contend that the *reduction* of thought experiments to arguments does nothing to demystify them, and indeed only serves to mystify arguments.

The basic problem of thought experiments is typically couched in Kuhn's terms of justification - how is it that beliefs may be formed with no new 'input' from the world? Familiar worries about the *a priori* and the modal status of conceivable things enter the scene here - if ideas about Cartesian demons aren't generated by sense data, what does it mean for us to imagine them? If our imagination produces a new belief, what justifies that belief? The framing of this debate in terms of evidence and justification is inherited from the putative experimental (and experiential) nature of thought experiments. It is deeply tied to the idea that thought experiments are about mental phenomena. If that phenomena is produced by an untrustworthy source, like a daydream, then it is unjustified. Thought experiments seem to be produced in such a way, thus their

[44] And indeed, any given thought experiment in my Model Engineering account can be reconstructed as an argument – just an argument with a very limited set of possible conclusions that do not include conclusions about the world.

conclusions are unjustified. Norton's argument view is framed against this kind of worry. Fear not, claims he. If thought experiments are merely arguments, then they are no more worrying than arguments. The phenomenology of the thought experiment is fundamentally just heuristic set-dressing, and the real justificatory core are premises rooted in experience, like any other argument. Thus, Norton's account seems to neatly wrap up the problem that the imaginative component of thought experiments poses for the justification of their conclusions.

I argue that this framing of the problem of thought experiments is actually somewhat misleading. The problem, so presented, is rooted in an epistemology built around questions of justification and truth. Norton's account implicitly reframes the problem in terms of language and meaning. This does not solve the problem, it merely relocates it. Reframing the problem of thought experiments as a problem about language reveals that the real issue involved is one of meaning and reference, not of evidence and justification. The wildness of the phenomena of thought experiments reasserts itself in this new framing. Demons, infinite empty spaces, swampmen, and the other weird and wonderful characters that populate thought experiments admit of no obvious real-world referents for which experience could give us insight. Norton's argument view has no obvious way to tame the epistemology of thought experiments in light of this new problem.

If thought experiments as arguments have the same epistemic characteristics as arguments, then it is important to understand what the epistemic characteristics of arguments are. The quintessential example is a simple Barbara syllogism, reprinted in every introductory logic textbook:

Socrates is a man

All men are mortal

Therefore, Socrates is mortal.

This short and rather convincing syllogism is about Socrates, who is a man. The truth of the conclusion that Socrates is mortal is guaranteed by the truth of the premises, which seem fairly plausible in themselves. One scarcely has to verify that Socrates is, in fact, dead. The meaning of this syllogism and its pertinence to the world is clear. Here's another syllogism:

The Ogopogo is a snake

All snakes are legless

Therefore, the Ogopogo is legless.

The Ogopogo syllogism is also a perfectly valid Barbara syllogism (if you don't like the 'legless' predicate, you can easily reformulate the syllogism into a Camestres of equal validity). Just as you did not need to read the Phaedo to convince yourself that Socrates was mortal after reading the above Socrates syllogism, you should need no further convincing that the Ogopogo has no legs. But the Ogopogo syllogism gives us a problem that Socrates does not - there is no Ogopogo. The Ogopogo is like Russell's Present King of France - a fictional entity. No more can premises about the Ogopogo be grounded in experience than premises about the hirsuteness of a putative Louis XXIX.

It seems intuitive to us that the Ogopogo syllogism is *about* a fictional entity and the Socrates syllogism is *about* a real entity. Both seem adequate to the task of telling us something we implicitly knew about their referents, but the way they do so is not the same. What makes Socrates a man is a fact about the world – one we know from historical accounts[45]. But what makes the Ogopogo a snake is a myth. When we say the Ogopogo is a snake, we are telling a story, not reporting a fact – a

[45] This even may be too far back in the historical record for good empiricist evidence-gathering. If you prefer, substitute 'Socrates' for 'Alex Trebeck' or some other epistemically available former mortal.

story made true only by convention. If every British Columbian cryptozoologist decided to think of the Ogopogo as a plesiosaur instead, the premise 'The Ogopogo is a snake' would be false and the premise 'The Ogopogo is a plesiosaur' true instead. We can easily point to what grounds the truth of these premises – historical facts and Canadian local legend, respectively – but those grounds are not the same.

What should we say, though, about the entities that appear in the premises of Norton's reconstructed thought experiments? Do they have the evidential status of Socrates or the Ogopogo? If the former, they can tell us something about the world. If not, they can only tell us about the fancies of our own minds.

There is no one answer to this question. Some of the referents of TEs are clearly real and accessible to empirical evidence. Some are conventional constructs like the Ogopogo. Some are idealizations of or abstractions from real entities and properties. Many have ambiguous status. Brown's account featured spooky imaginary justification, but the grounds of that justification were all decidedly real – real Platonic forms, with real instantiation in the world. This is not true of the Argument View. The Argument View has no resources at its disposal to deal with the actual epistemic problems of using imaginary entities to derive real results. As such, the view is left in a bind: it can either admit only premises that are empirically assessable (and thus reduce the set of valid thought experiments almost to nothing and exclude nearly all the celebrated examples) or it can admit premises that refer to fictional entities and give up its claim to solid justification of worldly claims. Neither, I suspect, is an attractive option.

2.3.4 Imagining the Ogopogo

The early 1990s saw the near-simultaneous (but independent) publication of two papers that made a similar point: Nancy Nersessian's 'In the Theoretician's Laboratory' (Nersessian 1992) and Nenad Miščević's 'Mental Models and Thought Experiments' (Miščević 1992). Both authors argue that there are some obvious similarities between thought experiments and mental models in science, and that these similarities can ground an account on which thought experiments simply are mental models. This account folds the epistemology of thought experiments into the epistemology of models in general.

Nersessian and Miščević both claim that the traditional epistemic wrangles over thought experiments have mischaracterized the actual mechanism by which thought experiments function. Instead, they claim that thought experiments are mental models of their target worldly phenomena and can be used to reason about the world in the same way. Thought experiments stand in some sort of relation to the natural world such that inferences made with them are thereby also inferences about that world and its contents. Those inferences are reliable because they are grounded in the human faculty of geometric-spatial cognition, a mental system that is itself reliable (though not infallible). Thus, there is no difficult question of how thought experiments reach the world, or at least no new one. The debate over how scientific models can show us anything about the world was covered at length in the previous chapter, but its details are irrelevant here: the point is that nobody doubts that they do. If thought experiments work like models, it seems, we need not fear for the justification of their conclusions, nor posit any strange and spooky mechanisms to justify our access

to them. Nersessian especially plays up the ordinary reliability of human visual-spatial mental reasoning[46] to justify this claim.

This account has many appealing features[47]. It provides a negative answer to Kuhn's problem about the novelty of the products of thought experiments by allowing that the answers thought experiments give were already present within the model. It provides an answer to the Ogopogo problem by appeal to the reliable-seeming human capacity for visual and spatial reasoning (Nersessian relates, for instance, the classic example of determining from memory how many windows are in your house by imagining walking from room to room). It seems to capture the abstractness of thought experiments without any extra metaphysical commitments, since mental models are typically abstracted off reality. The account pushes the epistemological question of the warrant for inferences from thought experiments to the world onto an account of the warrant for inferences from models to the world. From there, one can build an integrated account of both models and thought experiments together, which seems to solve two problems at once[48]. I would also posit that the association between models and thought experiments is a very natural one – the objects of thought experiments, the images and machines of thought, are clearly more akin to mental models than they are to propositions. It is for these reasons that my own account of thought experiments centrally features models - but as the objects of thought experiments, not thought experiments themselves.

[46] This reliance on visual-spatial reasoning does imply that Nersessian's account can only make sense of thought experiments that have a visual imaginative component. I don't think that's a necessary feature of thought experiments. However, it is true that visualizable thought experiments make up nearly all of the canonical set of thought experiments considered in the literature, so this domain restriction doesn't run afoul of the desiderata I defined above.
[47] Of the going accounts in the literature, it is the one that is closest to my own.
[48] As Nancy Nersessian does in her book on the subject (Nersessian 2008).

In this section I am going to make that picky and small-seeming distinction, and then I am going to try to convince you to make that distinction too. Thought experiments are not models. It is my view that holding this distinction makes the epistemology of models and thought experiments much clearer than it would otherwise be. I hold that thought experiments *feature* models, but in the same way that a laboratory experiment *features* a hypothesis. A hypothesis is a vital component of any experiment, but we would be missing something crucial about the epistemology of that experiment if we did not mark a distinction between the two.

An easy way to show the import of this distinction is to analyze the success and failure conditions of thought experiments and models, respectively. The success and failure conditions of thought experiments and models are not the same. These conditions are again analogous to the conditions of success for laboratory experiments and hypotheses respectively. A great experiment can be great because it truthfully showed the experimenter that its hypothesis was wrong. A bad experiment can be bad because it failed to reveal the truth of its hypothesis. Characterizing thought experiments as *tests* of models allows us to make this distinction but characterizing them as models simpliciter does not.

The history of thought experiments is a history full of glorious failure – that is, failure to produce imagined results that sync up to the world. Even if the mental mechanism by which thought experiments are carried out is the same ordinary visualization that allows us to turn down the correct streets on a walk or reach for the right shelf in the kitchen, the things that are being visualized are not so ordinary. After all, our ordinary spatial reasoning typically does not generate anomalies outside of the context of a dream, and as Thomas Kuhn pointed out, thought experiments very often generate anomalies – devastating problems for the theories in whose language the thought experiment is described (Kuhn 1977). The most famous of all the scientific

thought experiments, Galileo's thought experiment on falling bodies from the Two New Sciences, is one such famous source of anomaly. The upshot of Galileo's thought experiment is the destruction of the Aristotelian modeling framework in whose language it was couched.

But that Galilean ideal is not the only way that failure can manifest in thought experiments. Consider a similarly famous case – Einstein, Podolsky, and Rosen's argument against the completeness of quantum mechanics without hidden variables. EPR were clearly attempting to make an argument of the same sort as Galileo – that quantum mechanics fails to give an appropriate answer when faced with a sensible-seeming question. The fate of the EPR thought experiment, though, was not the same as that of Galileo's Falling Bodies. John Bell's reformulation of the thought experiment rendered it empirically testable, and Alain Aspect empirically tested it. The quantum formalism that EPR challenged prevailed, and EPR did not.

All this is clear with the benefit of hindsight. Empirical tests in the end settled the question of how to interpret the seeming failure of the thought experiment. However, even if there is no empirical evidence that could be called upon to settle the issue, the two alternatives described are still clearly different states of affairs[49]. In one case, the seeming conclusion of the thought experiment was true, in the other, it was false. For Galileo, the success of the thought experiment was the failure of the model it featured. For Einstein, Podolsky, and Rosen, the failure of the thought experiment was the success of the model. These, I claim, are the two different ways in which failure can arise in our thought experiments. We should demand that an account of thought

[49] This problem has a kind of analogue in laboratory experiments – the Experimenter's Regress. Harry Collins's regress is an epistemic problem that comes out of instrumentation. If one specific kind of procedure can be used to detect some heretofore unknown phenomenon, there may be no way to settle the question of whether the detector works or not. Collins' illustration of this case, the controversy surrounding Weber's gravitational wave detector, is such a story. If the machine returns a negative detection event, we are still left with two possible states of affairs: either the machine works and gravitational waves have not been detected, or the machine doesn't work and the gravitational waves that are there have passed by without tripping the detector. Without some other source of evidence for or against the success of the machine or the existence of gravitational waves, this question cannot be settled empirically (Collins 1991).

experiments marks this distinction. Accounts that equate thought experiments and models cannot do so. This is a version of the Ogopogo Problem that I defined above. The human faculty of mental modeling is very powerful, and I'd even allow that it's very reliable. But it is not infallible. Some of the failures in thought experiments are not failures of scientific theory or model – some are failures of thought experiments *as such*. We must hold those failures separate from the ways in which thought experiments reveal the failures of our model, or we will not know where we stand. We would be making the same mistake we would make by identifying an experiment with its hypothesis – for if experiment and hypothesis are one, what are we to think when we cannot replicate?

The way my account resolves this discrepancy is by separating the goals of thought experiments from the goals of models by giving up the claim that thought experiments have anything to say about the world. If models have the goal of describing the world (or providing opportunities for surrogative reasoning about it) and thought experiments have the goal of testing models, there is no conceptual confusion over the two kinds of failure in the foregoing examples. We may not have adequate empirical evidence to determine what state we are in following a thought experiment, but there's no question that the two states are different. Modeling accounts that hold to the claim that thought experiments can tell us about the world cannot hold this distinction.

2.4 Conclusions

In the previous chapter I developed an account of thought experiments out of considerations of scientific models and scientific change. In this chapter, I defined that account against the backdrop of other contemporary accounts of thought experiments. I showed that my Model Engineering account makes a crucial claim that other accounts in the literature do not – that thought experiments are non-evidential – and that in so doing, the account becomes invulnerable to two closely related challenges for accounts of thought experiments. The first challenge is Thomas

2.5.1 Why do some fields such as physics and philosophy have so many thought experiments (which are hugely important to those fields), while others such as chemistry have so few or none at all?

[50]

2.5.2 Literary fictions (novels, plays, movies) have a narrative structure similar to a thought experiment and they often teach us lessons of the same kind. Is this similarity only superficial or does it run deep?

One subtlety of my account is that thought experiments are processes, not entities. A thought experiment is a procedure a person undergoes in their brain using their concepts and models, not a thing written on a page. The words and images that record thought experiments also cause new people to try them out, but they are not the thought experiments themselves.

With that established, I don't think there's anything wrong with allowing that literary works can give rise to thought experiments. Catherine Elgin (Elgin 2014) has argued that fiction that encourages the reader to engage empathetically with its characters, like The Adventures of Huckleberry Finn, can aid a person's future moral reasoning in a manner similar to that of a moral thought experiment. My account does not necessarily cover moral thought experiments (as I will discuss below), but I do not think that rules out literary fiction from thought experiment on my account. Science fiction in particular is often written to cause us to push upon concepts we take for granted, like the nature of life and sentience. These stories can definitely give rise to thought experiments that accord with the way I have accounted for them. Looking at the fringes of our concepts in the way that science fiction encourages us to do exposes their open texture and can prompt us to reformulate them. However, it is also possible to read any piece of science fiction without performing a thought experiment if one reads it very literally. Moreover, if a reader's conceptual apparatus is sufficiently different from that of the author of a story, that story can revolutionize the concepts of that reader even if it would have seemed very ordinary to the author that wrote it. That is why it is crucial to maintain the distinction between the doing of the thought experiment and the record of it upon a page.

2.5.3 Does culture and background [of the people performing thought experiments] matter?

In my account of thought experiments, thought experiments are performed upon the modeling structures in the thought experimenter's mind. Thus, it is only natural that the contents of the thought experimenter's conceptual scheme matters tremendously to the results of their thought experiment. Some authors, most notably (Machery 2004) have advanced this argument, alongside empirical evidence of variation in intuitive and conceptual structures, as a kind of defeater of thought experiments as useful methods in philosophy. I don't think it is necessary to go quite that far. I think that, as long as one does thought experiments in the firm understanding that they do not provide evidence for the truth of claims, there is no problem with continuing to use them. Each human plausibly has their very own set of concepts and models that is unique to their situation and experience in the world. Exploring how well-constituted those conceptual schemes are is still distinctly worth doing. So long as philosophers and scientists are appropriately clear about what enters into the thought experiments they are doing, there is no special problem of cultural variability and thought experiments.

2.5.4 The legitimacy of thought experiments might vary from field to field. Does it?

It is no secret that contemporary philosophical thought experiments have a worse reputation than their scientific cousins, especially those on the physics side of the family. As I mentioned earlier in this chapter, this in part derives from the fact that 'thought experiment' is a term that can be applied both forward and backwards. Many of the most celebrated thought experiments in the sciences were named so in retrospect, rather than in by their own authors. Critics of thought experiments in philosophy are typically criticizing current thought experiments that are so called by their authors – typically the target of this critique isn't Descartes. So, we can read a bit of recency bias into this critique.

However, there is more to the critique than just the claim that a lot of contemporary philosophical thought experiments aren't very good. Philosophical thought experiments are accused of merely pumping intuitions the author already has about the subject in question, rather than providing any new reason to believe the claim in question. I think that is true, insofar as I think thought experiments do not, in general, give one a reason to believe in the truth of their result. However, it seems like the results of thought experiments in philosophy are more ephemeral and less useful than those in the sciences. There are famous thought experiments in philosophy, of course – Gettier cases are nearly a whole subfield unto themselves – but the published responses to them in the philosophical community are as negative as positive. The results of philosophical thought experiments for the conceptual schemes they test seem less substantial than their scientific counterparts, and less durable.

To mount a full explanation of this phenomenon would first require me to establish more firmly that it exists, and that a comparable instability in science does not. The literature analysis necessary to determine whether that is true is beyond the scope of this project. However, assuming for a moment that it is, here is how I would explain it. As I have laid out in my account, thought experiments are procedures that simulate the use of a model in a particular context to show whether or not the model can make sense of that context. Philosophical models typically have a major disadvantage as compared to scientific models for this sort of testing – the lack of mathematics. Scientific and mathematical thought experiments are nearly always couched in mathematical terms, and when contradictions appear in mathematics, it is typically not hard to see them. Philosophical thought experiments, on the other hand, are typically based in natural language. Natural language contradictions are much slipperier and much easier to iron away with a felicitous choice of words. So, two different instances of a philosophical thought experiment might seem very different, because the language they use is intrinsically less precise. Kathleen Wilkes makes a version of this

critique, specifically about the status of ethical and metaphysical thought experiments in (Wilkes 1988). Unlike Wilkes, I do not place the blame on the existence of natural kinds within scientific thought experiments and the non-existence of the same in philosophy. Since my account does not allow thought experiments to find the truth, the presence or absence (or, indeed, existence) of natural kinds is not relevant.

There is one more point I wish to briefly address – the status of moral thought experiments. It seems plausible to me that thought experiments in moral philosophy are of a different kind than thought experiments in other fields, due to the specific role that moral judgement plays within them. Ethical thought experiments often conjure a situation and then just ask the thought experimenter to render ethical judgement upon it. It is plausible to me that value judgements are of a different kind than the judgements of applicability that are central to my account. Thus, for the moment at least, I would like to leave the question of whether my account applies to moral thought experiments, such as the infamous trolley problem, open.

3.0. Lotto 1877

Boltzmann, Lakatos, and Model Engineering

> "*Certainly, therefore, Hertz is right when he says: "The rigour of science requires, that we distinguish well the undraped figure of nature itself from the gay-coloured vesture with which we clothe it at our pleasure". But I think this predilection for nudity would be carried too far if we were to forgo every hypothesis. Only we must not demand too much from our hypotheses."*
>
> - Ludwig Boltzmann, "On Certain Questions in the Theory of Gases", 1895. (Boltzmann 1895)

In the previous section I argued that thought experiments are fundamentally tools of construction and destruction, not of proof and disproof. Thought experiments test questions of model engineering – whether a model is well or poorly constructed, whether it can perform such-and-such functions, independent of its truth (or lack thereof) of the world it purports to describe. However, proof and disproof in science go along with construction and destruction. The destruction of a modeling structure takes all the proofs it provided with it, and it is impossible to prove any claim with no conceptual structure in place at all. The really revelatory models, the ones that change the world, are the ones that construct a new way of representing a phenomenon that immediately bears fruit by answering some sort of question. Traditional accounts of thought experiments blur these two processes together into one. I claim that if the two processes are held separate from each other it becomes much clearer where the warrant for the later claims comes from, and how much weight these claims have. This separation solves Kuhn's problem and the Ogopogo problem[51] in one fell swoop: it makes it clear where the justification of the proofs made on the basis of the

[51] As described in the previous chapter

representation come from (solving Kuhn's problem) and allows that not all imaginings give rise to proofs (solving the Ogopogo problem).

In this chapter, I will demonstrate the process of construction and proof in two toy examples of thought experiments in applied (and very applied) mathematics, and then move into a detailed tale of a real example from the history of physics – Boltzmann's lottery. I will argue that the process in both the toy and real examples is the same: building and testing representations using thought experiments. In all the cases, the choices made in building the models have nothing to do with the world of experience; but in all cases, the choices constrain the possible results the model can deliver. This constraint is what makes it possible for models to answer questions, but it also restricts the domain of the phenomena to which they can be applied. Model building is a process of give-and-take in which the twin demands of generality and power are traded off against each other until a satisfactory model is found. This dialectical process is very similar to the process of proof development described by Imre Lakatos in Proofs and Refutations.

I will contend that the role of thought experiments in science is typically misunderstood, and Boltzmann's Lottery is a case in which that misunderstanding makes a difference. Most accounts of thought experiments in the literature are *evidential*: they present the function of thought experiments as establishing or debunking certain facts. I argue that thought experiments are non-evidential, and that their only proper function in science is model engineering. Thought experiments allow scientists to work through the application of modeling techniques to particular systems. Not all models work for all applications, not all work well together, and not all work well enough to be of any use to science. In order to begin to use a new modeling strategy, scientists must first establish that the strategy is actually good for the purpose at hand. A good model must be coherent itself, be coherent

with the goals to which it is put, and coherent with the other models with which it must be used in order to fulfil that function. Thought experiments test these coherences.

It might be objected that any non-evidential account of thought experiments must account for the fact that thought experiments often appear to be evidential. Thought experiments seem to perform many functions, and proponents of other accounts have a stable of canonical examples of thought experiments functioning in the way they claim to deploy when examples are needed. I think I can explain away all these examples, as I have discussed elsewhere. It will perhaps be more elucidating for me to provide an example of my own. If I am right about thought experiments, paradigmatic thought experiments proceed like the discussion of Lakatos' imaginary mathematics students grappling with the Eulerian lemma – working, example in hand, through the consequences of the way they have represented their world, hunting for contradictions and incongruencies, proposing and building new applications and extensions, and finally making a decision about whether the representation succeeds or fails. This process typically starts from a naïve, pre-theoretical version of the representation, then works through the applications of the representation – proofs – until they begin to function properly. Boltzmann's lottery is a case of this kind. In the lottery paper, Boltzmann lays out the entire process of building his probabilistic representation of a gas, from the initial idea to the full model. Some of the ways of probabilistically representing the gas turn out to be incoherent. Some are just too simple and abstract to serve Boltzmann's purposes. All encode hidden lemmas that implicitly restrict the domain of application of the model. And the final representation, the one that Boltzmann only reaches after thought experimental trial after thought experimental error, allows Boltzmann to do just what he wanted: construct the quantity Ω that he needed to explain the approach of a gas to equilibrium.

The primary example discussed within this chapter is an example of a positive thought experiment – a thought experiment that demonstrates the coherence of a model rather than an incoherence. Thought experiments that demonstrate coherence in a model are no rarer than thought experiments that demonstrate incoherence, but they are less prominent. If a thought experiment shows that a well-used model falls apart under some novel circumstances, that failure is very notable. If a thought experiment merely shows that a well-used model continues to work as expected, that is less so. Positive thought experiments, like the Boltzmann example we are soon to consider, are typically only interesting in the context of the birth of a new way of representing a given phenomenon, when the question of whether it holds together at all is still of significant interest. As I will show later, Boltzmann's Lottery has been of little interest to physics and philosophy of physics since the acceptance of the modeling framework it tests. It is only here, within an exploration of the context of the birth of Boltzmann's framework, where its core representational idea needs to be honed, refined, and shown to hold, that the thought experiment really matters. So, let us see how thought experiments build.

First, however, let us get a grip on the way in which thought experiments refine models with a few toy examples. Each of the toy examples demonstrates how a thought experiment can bring along with it the construction of a model of some phenomenon, which gives both inferential power and limitations.

3.1 Toy Examples

3.1.1: Toy Example 1: Turning the Tables

Mathematicians, when left alone for a sufficiently significant span of time, begin to try to solve problems. Here's one such problem, which first appeared as a mathematical game in the 1970s but has periodically re-emerged in recreational and non-recreational mathematics journals ever since.

The patio at CERN is made of rough, uneven paving stones. Tables placed on such a surface have a bad habit of resting on just three of their four legs. This means that the slightest pressure on the side of the raised leg will cause the table to rock back and forth, spilling any mathematically necessary coffee that happens to be upon it. This is undesirable. Can applied mathematics help?

The physicist André Martin, one of the unfortunate CERN scientists losing coffee to the terrible tables day after day, published a brief informal proof of a solution to the problem (Martin 2007). He proved, first informally and later more formally, that even on uneven ground, there is always a way to place the table such that all four legs are resting stably on the ground (and hence, protecting the coffee from sudden changes in elevation). The informal proof goes like this:

Consider a symmetrical square table with four legs of equal length. It's sitting on ground that is uneven but not discontinuously uneven. Let's label the legs clockwise 1, 2, 3, and 4 and say, without loss of generality, that leg 4 is off the ground. Now imagine continuously rotating the table clockwise by 90° so that leg 4 is in the previous place of leg 1, leg 3 in place of 4, and so forth. Since the table is square, the new position of the table must look just like the old one – with three legs (4, 1, and 2) on the ground and one leg, 3, off the ground. That means that at some point in the rotation, leg 4 must have touched down from its elevated position and leg 3 must have lifted off. But the rotation was continuous – so the point at which must 4 touched down and 3 touched off must have been somewhere along that 90° rotation. At that point, all four legs were in contact with the ground. Therefore, there must be a way to set the table on the uneven surface so that all four legs are sitting on the ground – because we just found it. QED!

As Martin hastens to point out, this proof is not sufficiently rigorous to satisfy a mathematician. For one, establishing the existence and continuity of the rotation of the table is not a trivial exercise. For another, the ground can be bumpy but mustn't be too bumpy. There are several

different proof attempts at general versions of the table leg theorem, all of which suggest different maximum slopes for the ground and different generalities of shapes for the table. Martin's rigorous proof represents the table feet as a set of four points on a sphere and renders the conclusion that the maximum slope over the surface is no greater than 15°. Baritompa et al., in their 2018 paper on the table-turning theorem represent what they call a 'real table' as a rectangle with four line segments of equal length connected to its corners at right angles, and end up finding that any table with legs longer than $\frac{1}{\sqrt{(1+r^2)}}$ (where r is the ratio of the long and short sides of the rectangle) will have a point at which it balances on the ground, so long as the ground is Lipschitz continuous with a Lipschitz constant no greater than $1/\sqrt{2}$[52]. The two rigorous proofs' slightly different constructions give rise to slightly different results, even though the underlying principles (the non-rigorous proof sketch I gave above) are the same (Baritompa et al. 2018).

Both papers, however, are careful to note that the mathematical objects their proofs rotate are not tables – they are mathematical objects. The legs of physical tables have some thickness, the tiles of real patios have some friction and discontinuity, the motions of real mathematicians trying to rotate their tables into stability are not smooth. The way of converting the familiar physical action of rotating a table into a set of mathematical objects precisely-defined enough to prove something about is non-unique, and the way you do it matters. Different ways of mathematically constructing the same referent give different results.

[52] Which does correspond to the same maximum angle that Martin finds in the case of square tables, ≈35.26°. However, the Lipschitz continuity required for the Baritompa et al. proof is a stronger condition than the simple continuity required in the Martin proof.

3.1.2: Toy Example 2: The Whole Nine Yards

The previous example was from applied mathematics, specifically the applied mathematics of drinking coffee on a patio. The next is a practical problem from garment sewing[53].

You, like any tailor of good taste, have decided to make yourself a pleated skirt. Let your waist measurement be W. You want the skirt to be pleated all the way around, and you don't want any of the pleats to overlap. What length of fabric do you need to make your skirt? At first, it seems as if I have not provided enough information to solve this problem. I haven't, for instance, specified the size or number of the pleats. But this intuition is misleading – the problem already contains all the information required. All that is needed is to think about what a pleat of fabric is.

Consider the cross section of a pleat. Fabric in a pleat is folded lengthwise, then back again in a Z shape, then pressed down flat. Thus, at every point of the pleat, there are three layers of fabric. So, no matter how many or few pleats you put in your skirt, and how big or small they are, if the skirt is continuously pleated all the way around and none of the pleats overlap, then there will be three layers of fabric at every point around the waist of your skirt. Your skirt measures W at the waist, and at each point along W there are three layers, so the total length of fabric in the skirt must be 3W. QED.

Counterintuitively, the size and number of the pleats is totally irrelevant to the fabric consumption of the skirt so long as the two conditions, continuity and non-overlappingness, are met. This result is totally general to skirts meeting those conditions. However, like in the previous example, those conditions have served to construct the phenomenon of which they are a proof. An

[53] To the best of my knowledge, I am the first to propose this general solution to the pleat problem. However, the idea that a wide skirt should have a circumference of three times the waist is a commonly cited 'rule' in sewing communities both for pleated and gathered skirts. The Z-shaped cross-section of a pleat is also quite obvious to the eye when one is sewing, so I do not doubt that this general feature of pleats has been noted before.

incredulous reader who recalls the folk etymology of the expression 'the whole nine yards' (purportedly a reference to the nine yards of fabric contained within a traditionally manufactured Scottish kilt) will have done some quick math and noted that, according to my proof, there should only be nine yards of fabric on the kilts of people who measure 3 yards around, which is true of only some small proportion of kilt-wearers. And indeed, the conditions necessary to give us the neat solution to the puzzle are special ones. In constructing this proof, I moved from the general case of 'skirt with pleats' to a very specific kind of pleated skirt, under relevantly idealized conditions. For, of course, it is not quite true that the amount of fabric needed to make a pleat of one inch in length is three inches – some very slight allowance must be included for 'turn of cloth' – the amount of extra fabric needed for the turn of the fold, which is more for a thick fabric and less for a thin one. On top of that, most pleated skirts, including traditional kilts, do not meet the two conditions, non-overlappingness and continuity. These two features, the limitation of scope and the idealization of the subject matter, are what allowed me to make such a clean and simple proof of the fabric consumption of the skirt.

In both the toy cases so far considered, the actions taken to provide a representation of the system that can be used for the proof in question implicitly restricted the scope and descriptive accuracy of those proofs. This is precisely the process that Lakatos describes in *Proofs and Refutations*, just applied to a practical, rather than theoretical, case[54]. The tighter the representational grip of science gets on a particular domain, the more of the domain slips through its fingers. It is not an accidental feature of science that power trades off against generality: the power of a representation *comes from* the specificity of its construction. And eventually, as Lakatos shows with 'polyhedron' and

[54] An application that Lakatos also recognizes. Indeed, he claims that the methods of P&R are natural scientific methods applied to mathematics. I don't think that's quite right, but the sympathy between Lakatos' story of mathematical development and the story I will tell of scientific model development is obvious. I think that the sympathy is a result of the mathematization of science, though, rather than the application of natural scientific methods in mathematics, as Lakatos argues (Lakatos, Worrall, and Zahar 1976).

the Eulerian lemma, a structure that allows for perfectly tight and exceptionless derivations becomes no more (and no less) than a definition. By constructing 'table' in a certain way, we were able to learn something powerful about tables. By constructing 'pleated skirt' in a certain way, we were able to learn something powerful about pleated skirts. But in all three cases, our construction required us to eliminate certain members of the natural language extension of the terms 'polyhedron', 'table', and 'pleated skirt', and to fix all three to a particular level of idealization. The Small Stellated Dodecahedron, three-legged tables, and traditionally manufactured Scottish kilts have no place in the newly constructed regime of representational power.

These examples are toys, suitable more for recreational mathematics magazines than for the pages of *Nature*. But the principles that make these toy cases interesting – the trade-off between generality and power, the way that different approaches give rise to representations with different properties – are all real factors that have an effect on scientific modeling. The ways we choose to build models often feel *natural*, but the choices are never *forced*. We choose them every time, and our choices have consequences for the power of our representations that go beyond any question of model-world fit. This process is just as necessary in real model-building as it is in recreational mathematics. Scientists in the process of developing new models must decide how to mathematically represent the phenomena of interest, then test the consequences of doing so. If the process goes well, the scientist will find themselves with a new model that can answer the questions they wanted to answer. Let's now turn to a real example: Ludwig Boltzmann's lottery thought experiment in statistical mechanics.

3.2 Real Example: The Lottery Analogy

A large portion of Ludwig Boltzmann's seminal 1877 paper on the approach of a gas to equilibrium, "On the Relationship Between the Second Fundamental Theorem of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" is devoted to a carefully constructed analogy between a model of an atomic gas and a lottery machine. The lottery thought experiment is present in every section of the paper and is constantly modified alongside the model of the gas system itself. It is ubiquitous, central, and impossible to ignore. Yet, in august historical presentations of Boltzmann's achievement in the paper (Cercignani 1998; Uffink 2007; 2014) the lottery machine analogy either does not feature or receives only a passing mention. The argument of the 1877 paper has also since become a standard feature of thermodynamics textbooks. In these presentations, the lottery machine analogy again makes no appearance. It is by no means clear how such a central and striking feature of the 1877 paper could be as insignificant as to be eliminated from all subsequent rational reconstructions of Boltzmann's work.

This case is notable amongst thought experiments for a few reasons. First, it is difficult (if not impossible) to explain with other accounts of thought experiments. The obvious struggle on display in Boltzmann's paper does not cohere with the luminous 'aha' of a platonic account. The total absence of either empirical justification or argument for Boltzmann's construction makes an empiricist argument view unpalatable. And any view that considers the role of a thought experiment to be sounding the vast bay of possibility must contend with the manifest and obvious impossibility of every one of Boltzmann's lottery analogues. To get any real handle on what Boltzmann is doing when he asks us to imagine a lottery, we need to treat the thought experiment as a process of building a new conceptual structure, and then jumping on it a few times to prove that it can hold.

Boltzmann has given us the full walkthrough of a process that most texts only represent in part: the thought experiment complete. Let's see what that looks like.

First, I will walk through the lottery example in some detail. Then, I will discuss some features of its historical context that must inform our interpretation. Last, I'll discuss what this case shows us about the real role of thought experiments in science.

3.2.1 The Context of the 1877 paper[55]

The question that animated most of Boltzmann's thermodynamic work was the relationship between macroscopic and microscopic thermodynamics. The phenomena of macroscopic thermodynamics had, at the time of Boltzmann's writing, been well established for several decades. However, the scientific community had no microscopic account that cohered well with macroscopic thermodynamics, and indeed, the community was plagued by disagreement about what such an account should look like. Boltzmann was a staunch defender of atomism, and has become notorious for his early faith in the atomist project. However, Boltzmann was not what we would now think of as a scientific realist. Later in his life, Boltzmann championed what he called a '*Bildtheorie'* of science in which pictures, particularly simple and familiar pictures, are the fundamental ingredients and primary goals of a scientific theory. These *Bilder* were tools of the understanding – mental pictures that could allow scientists to get some grip on the phenomena they investigated (De Regt 1999). The extent to which Boltzmann had a well-formulated epistemology of science in mind before his turn to philosophy at the turn of the century is debated. Boltzmann himself claimed publicly to have read

[55] My summary of Boltzmann's argument will retain Boltzmann's own sometimes laboured terminology, since it is necessary paper to treat the steps as Boltzmann did in order to show the change in the lottery analogy. For a more succinct and terminologically clear version of the argument, see (Uffink 2014). Another well-known interpretation of the argument can be found in (Ehrenfest and Ehrenfest 2014).

very little philosophy and liked less of it, even as he was beginning to give public lectures on the subject (Boltzmann 2021). However, the same proto-pragmatist[56] tendencies found in the later Boltzmann's philosophical work provide a nice explanation of the relation between Boltzmann's great papers of 1872 and 1877, which are otherwise puzzlingly incompatible with each other.

Boltzmann had previously treated the task of building an account of microphysics to match the macroscopic theory of thermodynamics in his infamous H-Theorem paper of 1872. The 1877 paper is a second attempt at the same challenge. Boltzmann's goal in the 1877 paper is well-summarized by its title: it is a demonstration of a relationship between the approach to equilibrium guaranteed by the second law of thermodynamics and probability calculations. It has since come to be known as Boltzmann's 'combinatorial' argument, or the 'complexion-counting' approach to statistical mechanics. It is a fundamentally different argument than the one given in Boltzmann's controversial H-Theorem paper of 1872.

The 1877 paper is a retreat from the 1872 paper and uses none of the same material. The 1872 paper began with a combination of seemingly plausible dynamical assumptions about the bulk interactions of the molecules of a gas in order to derive an analytic proof of the second law of thermodynamics. If the assumptions hold, Boltzmann shows, a certain quantity (later denoted H but still at this point called E by Boltzmann) will necessarily monotonically decrease over time. By associating -E with entropy, Boltzmann appears to prove from only dynamical assumptions that entropy must increase over time, as macroscopic thermodynamics predicts.

However, both the dynamical assumptions that Boltzmann used to derive the H-Theorem and the scope of the result itself came under immediate scrutiny by the thermodynamic community

[56] I interpret Boltzmann's interest in scientific theories as useful metaphors and tools of thought as similar to modern pragmatist philosophy. See (Schmitt 2011) for another perspective, that treats Boltzmann's quiet epistemology as a case of Polyani-esque tacit knowledge.

of the time. Boltzmann's derivation seemed to pull a rabbit out of a hat – the assumptions all appeared to be plausible renditions of time-reversible Newtonian dynamics, but the result was not itself time-reversible. As Loschmidt pointed out in his 1876 response, Boltzmann's dynamical assumptions *must* have introduced irreversible dynamics into the gas system somewhere in the assumptions in order to derive the conclusion that entropy would monotonically increase, which is irreversible. So, Boltzmann's dynamical assumptions could not be straightforwardly those of Newtonian collisions between suitably idealized particles rattling around like billiard balls. The subsequent debate over this objection between Boltzmann, Loschmidt, Zermelo, and Culverwell, amongst others, would consume much of the subsequent two decades (H. R. Brown, Myrvold, and Uffink 2009). This prolonged and often perplexing debate may go some way towards explaining Boltzmann's decision in the 1877 paper to make a new argument with the troublesome dynamical assumptions to their barest minimum. This is where the lottery machines enter the picture.

Boltzmann does not argue for the analogy between the lottery construction and the motion of a gas, he merely states it:

> "It is clear that every single uniform state distribution which establishes itself after a certain time given a defined initial state is equally as probable as every single nonuniform state distribution, comparable to the situation in the game of Lotto where every single quintet is as improbable as the quintet 12345. The higher probability that the state distribution becomes uniform with time arises only because there are far more uniform state distributions" (Boltzmann et al., 2015, 1975)

Here Boltzmann is quoting his own paper of earlier the same year during the debate with Loschmidt. It is a remarkable demonstration of how simple the premise of the lottery paper is – and how central the analogy of the lottery is to it. Boltzmann's claim is just that if there are many more

states of equilibrium than nonequilibrium, all states are equiprobable, and a system is moving between states for long enough, then we should practically always expect the system to be in an equilibrium state once it has been given enough time to relax. The whole content of the paper is in determining the best way to quantify 'many more'.

Unlike the H-Theorem paper, the Lottery paper starts with almost no dynamical assumptions. Boltzmann's only stated assumption is that the molecules of the gas that he is modeling are capable of exchanging their kinetic energies (later in the paper, their directional velocities and momenta) by collisions. Uffink (2007) notes that there is a second concealed dynamical equation contained in Boltzmann's formula for the total energy of the system. Because the total energy of the system is expressed as a simple sum of the energies of all its component particles, the energy of the individual particles cannot depend on the states of the other particles – that is, there can be no interaction between them. This amounts to the assumption that the gas Boltzmann is modeling is an ideal gas. So, Boltzmann's lottery paper is not entirely free of dynamical equations, but the dynamical assumptions on the gas molecules seem to lack the dubious directionality of the H-theorem. Certainly, there is nothing as objectionable as the H-theorem's *Stoßzahlansatz* lurking in the wings of this proof. The simplicity and apparent generality of Boltzmann's dynamical assumptions suggest a problem, however. Irreversible dynamics cannot come out of reversible dynamics. That is the heart of the reversibility objection. The simple gas dynamics of the lottery paper do not include any element that obviously lead to irreversible dynamics.

Boltzmann gets away with this light touch in the gas dynamics by introducing the image that will be the topic of the rest of this paper: the lottery analogy. The dynamics of the lottery are not realistic (for instance, when the lottery becomes infinite, Boltzmann makes no attempt to understand

how an infinite urn containing infinitely many slips of paper would be possible) but they encode the core of Boltzmann's strategy: by stipulation, each slip of paper in the lottery has the same probability of being drawn as any other. The equiprobability of the lottery slips represents the equiprobability of the different complexions[57] of Boltzmann's gas model. However, the specific meaning of the elements of the lottery metaphor evolve alongside the model of the gas itself. Let us break down Boltzmann's metaphor in detail.

3.2.2 Playing Lotto 1877

Boltzmann's paper begins with a simple toy model of a gas and a simple lottery analogy. Both models evolve towards greater complexity throughout the paper as the simpler constructions are rejected. As the gas model becomes more complex, general, and realistic, the lottery model becomes stranger, more involved, and less realistic. In this section, I will walk through the stages of these parallel transformations in some detail. This treatment is not intended to be a full account of the derivation of the 1877 paper, which has been thoroughly discussed in other presentations of the history of thermodynamics. Instead, I will describe only the lottery machine metaphors and their interaction with the gas system as they arise.

[57] Boltzmann does not use the modern convention of dividing the state of the system into 'Macrostate' and 'Microstate'. Instead, he uses a threefold division of his own devising, which separates the microstate into two different levels of description. The 'State' of a gas corresponds to phenomenal thermodynamics and its observable quantities, like pressure, temperature, and volume. The 'State Distribution' is the next layer down. It is a statistical description of the microscopic properties of the gases - a census of how many particles have each given value of energy but agnostic about which particles have which. The lowest level of description is the 'complexion', which is the complete specification of the energy state of every individual molecule of the gas. The separation of the state distribution from the complexion is the key move that allows Boltzmann's strategy to work. I will retain Boltzmann's terms throughout. See Sharp and Matschinsky's preface to their 2015 translation of the 1877 paper for more detail (Boltzmann 2015).

3.2.2.i The Discrete Energy Lottery

Boltzmann begins with a simple model of a gas and a simple model of a lottery. The first gas system is a finite gas of *n* molecules, each of which can take on a kinetic energy value that is an integer multiple of some value ε. The total energy of the system (L) is a constant multiple of ε such that $\lambda\varepsilon = L$. Boltzmann starts with $n = 7$ and $\lambda=7$. Boltzmann then leads the reader through the simple combinatoric exercise of determining how many complexions (ways of distributing the 7 units of energy between the 7 molecules) correspond to each of the 15 possible state distributions (numbers of molecules with each given energy, like 0000007 or 0111112). Most state distributions have many complexions associated with them, the extremal cases 0000007 and 111111 have only one apiece. This follows only from classical combinatorics – no assumptions about how the molecules would arrive in one of these states have yet been made.

The crucial turn in the argument happens after Boltzmann has laid out the numbers of all the permutations (1978). After showing that there are only 7 possible complexions corresponding to the state distribution 0000007 (since the one molecule with a kinetic energy of 7ε could be any of them) he makes the jump from counting to probability by invoking the lottery analogy directly. This is the philosophical core of the paper, so I will quote it at length.

> "Denoting the sum of all possible complexions, 1716, by J then the **probability** of the first state distribution is 7/J, similarly the probability of the second state distribution is 42/J; the most probable state distribution is the tenth as its elements permit the greatest number of permutations. Hereon, we call the number of permutations the **relative likelihood** of the state distribution; this can be defined in a different way, which we next illustrate with a specific numerical example, since generalization is straightforward. **Suppose we have an urn containing an infinite number of paper slips**. On each slip is one of the numbers 0, 1, 2, 3, 4, 5, 6, 7; each number is on the same number of slips and has the same probability of being

picked. We now draw the first septet of slips, and note the numbers on them. This septet provides a sample state distribution with a kinetic energy of ε times the number written on the first slip for molecule 1, and so forth. We return the slips to the urn, and draw a second septet which gives us a second state distribution, etc. After we draw a very large number of septets, we reject all those for which the total does not equal 7. This still leaves a large number of septets. Since each number has the same probability of occurrence, and the same elements in a different order form different complexions, **each possible complexion will occur equally often**." ((Boltzmann 2015) 1978, emphasis mine)

This passage is the first occurrence of the concept of probability in the paper. Boltzmann simply defines the probability of a given state distribution as the number of complexions corresponding to it divided by the total number of complexions. This notion of probability does not fall out of Boltzmann's dynamical assumptions. It has nothing to do with dynamics. It does not correspond to any property of the gas. It is merely stipulated in.

Boltzmann then immediately supplements his definition of probability by introducing a lottery procedure that would produce the same probabilistic structure. According to Boltzmann it simply follows from the fact that there are as many slips with the number 2 on them in the (infinite) jar as there are with the number 5 on them that drawing a slip labeled 2 is just as probable as drawing a slip labeled 5. And certainly, any lottery in which it was not true would not be a lottery one would want to play. But the probabilistic character of the lottery system has no more dynamical underpinning than the gas system. For instance, as I will discuss in section 3, Boltzmann has not clarified whether the probability of drawing a given slip (and thus, the probability of a given state distribution) ought to be understood as objective, subjective, or neither. With the lottery, Boltzmann

has merely introduced us to a second system that has the features he wishes to claim are present in the gas system.

At this stage it is also worth noting that the lottery analogy is not separate from the gas that Boltzmann is modeling – it is embedded in it. When we draw a septet of tickets from the lottery urn, what we get is immediately identified with a complexion and state distribution of the gas. So, we can see that the lottery analogy is not merely providing an illustrative flourish to Boltzmann's notion of probability. The lottery analogy provides the descriptive content for Boltzmann's notion of probability.

3.2.2.ii The Continuous Energy Lottery

Boltzmann's next move is to generalize the formulae that were used to find the number of complexions for each state distribution to large numbers of particles and an infinite number of small energy units ε. The ceiling on the number of energy units ε must be infinite so that Boltzmann can send the size of the energy units to (almost) zero. Infinitesimal (but still discrete) energy units are the stand-in for a continuous kinetic energy value throughout the construction. Boltzmann sets the energy increment ε small enough that he can safely consider kinetic energies between x and x + ε to all be equal to each other. This is still a discrete partition of the supposedly continuous energy variable, just a very fine-grained one. However, it does require a modification in the set-up of Boltzmann's lottery.

The Discrete Energy Lottery was already an infinite lottery. It was infinite in the sense that it contained infinitely many slips, and equally many slips of each of the 8 kinds. Infinitely many slips of each kind meant that Boltzmann could assume that the probability of drawing each slip in turn remained equal no matter how many slips were drawn. However, Boltzmann could have obtained the same result simply by specifying that there were equal numbers of slips of each kind in the urn

and that each slip would be replaced after being drawn. That is, the infinity of the first lottery was not necessary to it. This is not true of the Continuous Energy Lottery. The Continuous Energy Lottery is infinite in two ways: it has infinitely many tickets of each type, and (countably[58]) infinitely many types of tickets. So, while the previous lottery was intuitively visualizable as a very large lottery (or a lottery with replacement, to which it is equivalent), this second lottery is not so easy to imagine[59]. But it is still different from the previous lottery in only this respect. Each of the infinitesimal 'steps' up the infinite energy scale is equiprobable, and the n-tet corresponding to a particular complexion is drawn from the urn in the same way as in the finite case.

However, after Boltzmann develops expressions for the probabilities of particular state distributions generated by this approach, he notes that the results of this approach do not, in fact, well-model a gas. In what Cercignani calls a Maxwell-inspired *coup de théatre* (Cercignani 1998) Boltzmann reveals that he has made an error in his construction: setting increments along the kinetic energy scale to be equiprobable to each other as Boltzmann has done in this lottery machine undercounts the energetic degrees of freedom in a three-dimensional gas. Instead, the method that Boltzmann has developed has the right number of degrees of freedom to count the complexions of a 'gas' made of discs in two-dimensions, or infinitely long cylinders. In order to actually get the quantity of interest, Boltzmann has to build a different lottery.

[58] Georg Cantor's first paper on the sizes of infinite sets was published in 1874, three years before the Lottery paper. However, if Boltzmann was aware of Cantor's work at this time, he does not demonstrate it in the 1877 paper. We can fairly say that this particular infinite quantity in Boltzmann's second lottery is countable because it is described as integer multiples of a certain very small quantity, and thus must be in a one-to-one correspondence with those integers. Boltzmann did study Cantor's set theoretic work carefully later in life and even lectured on him (Cercignani 1998; Tanaka 1999), but plausibly had not done so at this point.

[59] Indeed, there are conceptual problems with any physical realization of a lottery machine that selects one of a countable infinity of outcomes with equal probability. For details, see (Norton 2018b; 2020). Boltzmann does not here seem to be worried about any such problems.

3.2.2.iii The Continuous Velocity Lottery

It is important to note that each of Boltzmann's lottery machines so far has been one step of complexity removed from the previous one. I will speculate on the reasons for this in the analysis below. Boltzmann's third lottery machine continues the pattern. Boltzmann replaces the single infinite lottery machine marked with increments of kinetic energy with three lottery machines on the same model – one for each of the three components of velocity for each particle. As in the previous case, each increment along each velocity axis is equiprobable with all the others, and the axes go to infinity. Boltzmann has solved his modeling problem of the previous section. The lottery can now be said to well-represent the gas that he wants to model, and indeed, Boltzmann says just that:

> "To get the right distribution for the latter case [a gas] we must set up the initial distributions of paper slips in a different way. To this point we assumed that the number of paper slips labeled with kinetic energy values between 0 and ε is the same as those between ε and 2ε. As also for slips with kinetic energies between 2ε and 3ε, 3ε and 4ε, etc.
>
> Now, however, let us assume that the three velocity components along the three coordinate axes, rather than the kinetic energies, are written on the paper slips in the urn. The idea is the same: there are the same number of slips with u between 0 and ε, v between 0 and ζ, and w between 0 and η is the same… Here, u, v, and w have any magnitude and ε, ζ, and η are infinitesimal [finite] constants. With this one modification of the problem, we end up with the **actual state distribution** established in gas molecules" (Boltzmann 2015), 1989).

The third variation of the lottery model is success because it generates the 'right' or 'actual' distribution of states exhibited by gas molecules. This is the lottery that Boltzmann can use to build the concept of equiprobable states needed to build a generalized measure of permutability for a gas.

And indeed, it is at this point in the paper that the desired quantity Ω, the permutation number for a state distribution of a gas, appears for the first time.

Note that even this deep into the paper, Boltzmann is still identifying all probabilistic concepts with statements about the lottery, not the gas. For instance, when he makes the crucial identification between the most likely state distribution and thermal equilibrium, the most likely state distribution is still described as the "most likely sampling" from the lottery defined as above (ibid, 1990).

3.2.2.iv The Continuous Generalized Coordinate Lottery

Boltzmann has already reached the main result of the paper at this point (the derivation of the general triple-integral expression for Ω in an ideal gas – this is used to derive the first version of the $S = k \log \Omega$ relation in the final section) but he has one more puzzle to solve before he is done defining lotteries. The initial context of Loschmidt's reversibility objection to the H-theorem had been as a side-note in a response to a different Boltzmann paper – a 1875 paper about the action of a uniform field of force, such as gravity, on a gas in thermal equilibrium (Uffink 2014). The H-theorem predicted that the temperature, and therefore the kinetic energy of the gas would be uniform despite the gravitational field. Loschmidt contended that the molecules that were rising should lose kinetic energy by doing work against the gravitational field and thereby cool as they got higher, which would make the equilibrium state of the gas not also state of thermal equilibrium. Boltzmann considers his H-theorem approach to have already solved this problem. His new solution in the 1877 paper is undertaken merely in the name of generality. Indeed, Boltzmann gestures at this solution eventually being able to generalize not only to gases under external forces and multi-atomic gases (the obvious candidates) but also eventually to "any solid and liquid" ((Boltzmann 2015), 1993).

The fourth and final evolution of the same lottery metaphor Boltzmann gave us at the beginning is again, a single step removed from the previous lottery. The three urns filled with infinitely many slips divided into infinitely many incremental steps of velocity must multiply in number to account for the generalized coordinates of each molecule. Boltzmann still goes to the trouble of carefully laying out a procedure for randomly drawing values for each of the generalized coordinates for each of the molecules in turn. There are two variants of this lottery machine: first, a machine that gives the state distributions of arbitrarily many kinds of multiatomic gas; and second, a machine that gives the state distributions of gases under the influence of external forces. Though these lotteries are more complex, they do not stray from the general principles of the lotteries detailed above. For each of the generalized coordinates of a given molecule in a given system, there is an equal chance of it taking on any value of that coordinate because there is an equal number of slips in each urn corresponding to each infinitesimal value it could take on. Slips are drawn for each molecule in the gas, and any set of slips that does not add up to the total energy of the system is discarded. One state distribution will occur more often than any other when this procedure is followed, and that state distribution is defined as thermal equilibrium.

Despite the models of a gas becoming increasingly realistic and less idealized, Boltzmann is still committed to a model in which, "it is of course **entirely chance** that determines the state distributions for the gas molecules" (ibid, 1995). Even in the model of a gas that is being acted upon by external forces, a model that is intrinsically and inescapably dynamical, the question of what state the system is in is answered by the lottery, not by the dynamics.

3.2.2.v The Many-draw Lottery

Boltzmann introduces one more lottery machine in the 1877 paper. However, this lottery is presented as a variant that does not work as a representation of the gas. It the only lottery in the

paper that is not constructed in a stepwise fashion from the others. Boltzmann's text is very obscure in this section. The new lottery is presented out of the blue and does not appear to connect to anything else in the paper. Boltzmann introduces the Variant Lottery as a demonstration of "how general the concept of the most probable state distribution of a gas is," by defining it in a different way (ibid, 2001). The link between this multiple representational realizability and the generality of the relevant concept is typical of Boltzmann's *Bildtheorie* approach to model-building (De Regt 1999). However, Boltzmann is not satisfied by the results generated by this machine. The way of counting the permutation number of a given state distribution from this model does not give the 'correct' value for a gas.

Briefly, the Many-Draw Lottery approaches the same problem that the Continuous Energy Lottery does. But instead of an urn containing infinitely many slips with integer increments of energy which are drawn for each molecule in turn, the urn contains one ball for every molecule in the gas. The total kinetic energy of the gas is $L = \lambda\varepsilon$ for some small energy unit ε and integer λ. The total energy is 'doled out' by making λ draws from the urn. Each molecule is assigned one unit of energy for each time it is drawn. So, the probability of a given state distribution is the probability of drawing particular balls enough times to build that distribution after λ draws. However, after developing an expression for the probability of an arbitrary state using this method, Boltzmann asserts that it does not lead to thermal equilibrium and abandons the model.

It is not easy to see how this strange aside in the paper can be squared with the rest of the work done with lottery models throughout. Indeed, it is not even clear what Boltzmann thinks the upshot of this little digression is. I think it is best to consider the Variant Lottery a separate analogy from the four other versions of the lottery analogy that we have seen so far.

3.3 Discussion

3.3.1 Probability in 19th Century Physics

Boltzmann's Lottery is the only way he defines probability in this paper. So, in order to understand what the lottery analogy means in the context of the paper, we must understand how probability was used in the physics of the day.

It is perhaps not too strong to say that, of all the branches of mathematics, probability is the one whose relationship to the physical sciences is the most tortuous and ambiguous. The received view is that probability as a branch of mathematics was born already applied in the gambling games of the 1660s (Hacking 1975)[60]. Its applied beginnings may partially explain why its route into physical applications was as circuitous as it was. At the time of Boltzmann's writing in 1877, the role of probability in physics was fragmented between several different meanings. Even in Boltzmann's own work probabilistic concepts arise in a number of different ways. The aim of this section is not to settle any of the 19th century's debates over the meaning or meanings of probability – merely to demonstrate that the landscape in which Boltzmann was writing was one in which his use of probabilistic mathematics required additional clarification.

Broadly speaking, there were three dominant strands of probabilistic thinking coming out of the 18th century and into the 19th. There was a school of thought that made no distinction between objective and subjective probability, or moved between them freely (Gigerenzer et al. 1990), 16-18). The second strand, which emerged in the middle of the 19th century, imposed a firm distinction

[60] Though the absence of prior mathematical models of chance is now disputed. See (Norton forthcoming) for one such dispute.

between objective and subjective probability especially in the writing of Boole, Bertrand, and Mill (ibid, 36). The third strand, born of the 18th century movement towards statistics in governance and criminology and typically attributed to Quetelet, deemphasizes the explanation of chance and emphasizes the lawlike regularities that fall out of the law of large numbers. This sociological strand entered physics through analysis of measurement error in astronomy (ibid, 167-168) but was later the analogy of choice for Maxwell in his own discussions of the statistical character of gases (ibid, 62).

Much has been written about Boltzmann's ever-changing relationship to different concepts of probability. Most authors agree that Boltzmann's view evolved over time, but there is still substantial disagreement on when the evolution occurred and what the start and end points of that evolution were (Uffink 2007, 53; see Uffink 2014 for a summary of other positions). The question is not helped by Boltzmann's own tendency to read back into his own past work claims that are hard to find therein, as he did in the later debate over whether the H-theorem was exceptionless. It is also, prima facie, impeded by the oblique approach of the lottery paper itself. Boltzmann never clarifies whether we are to interpret the probabilities that the lottery machines in the 1877 paper give us as subjective or objective. Indeed, neither seem like a good fit for Boltzmann's machines and the relationship they bear to the gas they are supposed to model. Instead, I think it is best to understand Boltzmann's talk of probability in the 1877 paper as pure modeling strategy, rather than as a description of the physics of any real system. This brings Boltzmann's discussion here more in line with the *Bildtheorie* he would later espouse, and it makes better sense of the role of the lottery in the paper as a thought experimental prop for establishing and testing a new representation. Just like Martin's spherical table or my Z-shaped two dimensional pleat, the lottery gives Boltzmann the structure he needs to give content to his new statistical way of representing a gas.

First, if the probability for a gas to be in a particular state that we get out of Boltzmann's Lottery is objective, it seems like it must be wrong – and if we believe Bertrand, a wrong objective probability is worthless. The foundational assumption of Boltzmann's paper is that he is going to bypass the actual dynamics of a gas in the name of generality. The lottery machine certainly is not mimicking any of the dynamics of an actual gas. But as we have seen, Boltzmann freely trades between discussion of the gas and discussion of the lottery. The probabilities of the one are the probabilities of the other throughout – indeed, the lottery machine is how Boltzmann *defines* the probability of a state of the gas. It is hard to see how the attribution of equal probability to every complexion could be objectively true of any gas.

On the other hand, if the lottery is supposed to represent our subjective beliefs about the state of the gas, as the indifferent probability distribution would suggest, it becomes difficult to see how Boltzmann's paper actually proves anything about the approach of a gas to equilibrium. As Uffink notes, "the principle of insufficient reason, or any similar assumption, makes sense only from the view point that probability is a non-mechanical notion: it reflects our belief or information about a system." (ibid, 53) and there is little evidence elsewhere that this is Boltzmann's considered position. And indeed, it is not clear how a non-mechanical or informational picture of the probability of a state distribution would be sufficient to show why a given gas would approach equilibrium, since that is a mechanical explanandum. A subjective or epistemic account of the probabilities of the lottery might show that we ought to expect a given gas to approach equilibrium, or that most gases we will find will be in a state at or near equilibrium, but that is hardly a revelation. We know this already from experience.

However, I claim that these two short arguments are unsatisfying because they miss the point of the lottery analogy within Boltzmann's lottery paper. It is not a coincidence that most

authors who write on this paper do not mention the details of the lottery machines at all, or do so only in passing. It is also not a coincidence or a mistake that Boltzmann spends so much time developing the analogy. The reason is less literal and more interesting than a simple attribution of subjective or objective probabilities to states of a gas. In the next section, I will develop an alternative story that justifies the presence of the lottery machine analogy in the 1877 paper.

3.3.2 Justifying a Representation

Boltzmann's paper is about gases, but, at the risk of obviousness, it is not a gas. It is up to the readers of the paper to interpret the discussion of imaginary tiny hard spheres flying about as having anything in particular to do with a gas. It is easy to overstate the obviousness of modeling assumptions that are familiar to us. No modern reader, nor 19th century reader, would be confused at the language that Boltzmann uses to describe the molecular properties of a gas[61]. Boltzmann does not need to teach us how to ascribe properties like 'kinetic energy' to a gas or how those properties ought to be represented. The same is not true of ascribing probabilities – any probabilities at all – to the state distribution of a gas. The strategy of representing the properties of a gas by probabilistic structures itself needed to be justified by Boltzmann in order for his use of the assumption that complexions are equiprobable to be compelling. The lottery analogy serves the explicit purpose of justifying the use of that probabilistic representation. For 21st century readers used to the presence of probabilities in physics, that justification seems merely decorative, but in Boltzmann's own context it is as vital as any other part of the proof.

The question of what makes a scientific representation a good one is not new. 20th century philosophy of science has struggled over the status of non-literal presentations of scientific claims,

[61] Of course, many would have disagreed with it.

like idealizations, abstractions, and analogies, for decades[62]. However, the case of Boltzmann's Lottery would seem to show that the analysis of what makes a representation a good one is often less interesting than the question of how we learn and justify that a representation can be useful for the particular purpose at hand. It would be a stretch to say that there is a genuine mapping relation between Boltzmann's lottery and any real gas in the world – or if there is one, that it includes the indifferent distribution of probability over the complexions of a gas that is the core of the method. But with the benefit of hindsight, we can see that the statistical method of approaching gases that Boltzmann debuted in this paper is tremendously valuable as a way of defining and quantifying entropy. Even in the context of this paper, the proof of the proverbial pudding is in the eating: Boltzmann's lottery generates the quantity Ω, which can be integrated over to provide a value for entropy that matches the value derived using the macroscopic method familiar to Boltzmann since Clausius. The instructive question, then, is how Boltzmann can bring his readers along for the ride.

This, I claim, is the role of the lottery analogy within Boltzmann's paper: to justify the presence of probabilistic mathematical modeling strategies in the description of what is assumed to be a deterministic gas. At every place in the paper where Boltzmann has to re-define his complexion-counting strategy to adapt to a new context, he introduces a new lottery for which that counting strategy is natural and obvious. The step-wise progression of complexity within the lottery models carries the reader along with exclusively familiar examples. Even the early pitfall of dividing up the probability space by kinetic energy rather than by velocity components for a 3D gas is needed to provide the next rung of the ladder for the reader, since the two subsequent models both use that lottery as their base. The lottery analogy also spares Boltzmann the trouble of establishing either the origin of the probabilistic character of the system (how does it fall out of wholly reversible

[62] See Chapter 1 for more discussion of this debate.

dynamics?) or the meaning of the probabilities themselves (are they subjective or objective?). These conceptual difficulties are not thereby solved, of course. They persist to this day. But they persist as part of a much richer thermodynamics as a result of Boltzmann's work in this paper. Using an analogy system radically different from the scientific system in question to import desired mathematical features without having to thereby justify them is a common strategy in the history of scientific theorycrafting. Boltzmann's Lottery is an exceptionally good example of just how powerful a positive thought experiment can be.

3.4 Conclusion

In this paper, I have presented three examples of my account of thought experiments in action – two toy examples, and one historical example. The toy examples demonstrated the ways that decision making in model construction affect the power and usefulness of models independently of empirical evidence, and the ways in which the power of models comes from embracing limitations in scope. Then, I gave an account of the complicated and oft-ignored lottery metaphors that appear throughout Ludwig Boltzmann's 1877 paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium". I described how the thought experiment evolves alongside the derivation of the permutability measure in order to continue to justify it. I then argued that the reason Boltzmann's use of the analogy is so careful and thorough in the paper is that the purpose of the analogy is to justify a then-novel way of representing a gas with probabilities. The lottery is neither a subjective nor an objective probability. It is simply a modeling assumption. But just like the turning tables or the pleated skirts of the toy examples, Boltzmann's probabilistic construction of the gas puts implicit limits on that phenomena. The power of his new representation comes as at the expense of descriptive scope.

It would be convenient for the historian of science if the most influential publications in history were dry lists of propositions and their logical consequences. But the history of science is neither so simple nor so dull. An evocative image can change the direction of a field – even if the way it does so is by papering over where a theoretical edifice still has some missing bricks.

4.0 Reinventing the Wheel:
Paradoxes, Thought Experiments, and the Rota Aristotelica

"What God would set
Such incompatible truths loose
To struggle thus with one another?
Either could stand alone, but together
How can their contradictions be joined?
Or is there some way that they can get on,
That the human mind, enmeshed in flesh,
Cannot discern? That flame is covered,
And in the darkness the world's subtle
Connections are hidden."

- Boethius, *The Consolation of Philosophy* (Boethius 2009)

This introductory sentence is a lie.

Paradoxes are the first real introduction to philosophy for many students. The mind-bending nonsense of Lewis Carroll's paraconsistent worlds, the familiar trope of the science-fiction robot

smoking and sparking as it tries to compute an unparsable sentence, the Tortoise racing the Hare and winning… all are paradoxes, and all are, on some level, philosophical. The deep connection between the perplexity of paradox and philosophy makes it all the more surprising that few systematic attempts at understanding paradox *qua* paradox have been attempted[63]. One notable attempt, W.V.O. Quine's "Ways of Paradox" from 1966, provides a plausible explanation for this otherwise puzzling omission: the category of paradox is only an accidental association of unrelated concepts, with no shared essence. Quine thinks that paradoxes fall into three broad categories (the misconceived *falsidical paradoxes*, the merely surprising *veridical paradoxes*, and the revolutionary *antinomies*) and that individual paradoxes can change between the categories, but does not claim that the categories are connected by any overarching concept of paradox as such. If paradoxes have no shared essence there is no need, and indeed no way, to have an account of them *qua* paradoxes. Understanding why something is called a paradox would be a job for a historian, not a philosopher.

I think Quine's judgement of the disunity of the so-called paradoxes is premature, and in this paper I offer an alternative. Though Quine is right to note that the three kinds of paradox he identifies are distinct in their epistemic and pragmatic consequences, I argue that this difference obscures a deeper similarity in purpose. Paradoxes are thought experiments. Their purpose is to experiment upon the representational structures of thought. If paradoxes are understood in this way, the apparent disunity of the three kinds of paradox Quine identifies melts away – it becomes no more puzzling than the different possible outcomes of a laboratory experiment. Quine's error is to look at the experiment only once it has been performed, and thus sees only the positive, negative, and inconclusive results. By the time the results are plain to see, the common purpose of all these

[63] Many attempts to explain paradoxes focus on specific kinds of paradox, such as linguistic or mathematical paradoxes, to make the problem more tractable. For instance, the Stanford Encyclopaedia of Philosophy separates the topic of paradox into five separate articles (Logical paradoxes, The Sorites Paradox, Zeno's Paradox, Epistemic Paradoxes, and Fitch's Knowledge Paradox) (Cantini and Bruni 2021).

paradoxes has already been satisfied. Because I treat paradoxes as defined by this common model-testing purpose rather than as a mere set of unrelated canonical cases bearing the 'paradox' name, I can also better explain a puzzling feature of some canonical paradoxes that Quine raises in "Ways of Paradox": that they appear to shift their 'species' over time. In the second half of the paper, I walk through an example of this phenomenon more extreme than any Quine considers – the long and twisted history of the *Rota Aristotelica*, or Paradox of the Wheel. The Rota Aristotelica has undergone three major re-interpretations since its inception in antiquity, and each time it is reborn, its place in Quine's taxonomy changes.

However, in order to defend an account of paradoxes as thought experiments, I must overcome a general challenge: Ian Hacking's argument that thought experiments do not have lives of their own, and thus cannot shift and change over time[64]. If Hacking is right about thought experiments, then paradoxes must be something quite different. After all, Quine argues persuasively that paradoxes often change over time. At most two of Quine's, Hacking's, and my views can be true at once. In this paper I will argue for Quine's and mine.

Even if thought experiments have no life of their own, the Rota Aristotelica has at least a fascinating range of undeaths, as each new era of physics reanimates its corpse to serve them once again. In each of the three resurrections that I will discuss in this paper, the Rota Aristotelica was interpreted as a puzzle for a different domain, each time with a different solution. Each version of the paradox reveals a different kind of conceptual problem and each solution resolves it in a different way. I will argue that paradoxes work because they flow from attempts to *actually use* the representation in question for a purpose, and a different purpose each time. Were the wheel representation not actually used, the tensions of its use would not have been discovered. Each

[64] (Hacking 1992), "Do Thought Experiments have a Life of their own", a question that he immediately answers with 'no'.

reinvention of the wheel is a new use, thus each is a new tension. So, though the different solutions of the paradox conflict with each other, we need not conclude that one is right and the others wrong – instead, we merely need to ask which one tells us something we want to know.

4.1 What is a Paradox?

The canonical slate of paradoxes includes cases that produce results that are merely surprising consequences of our theories (eg. the Birthday Paradox, the Twin Paradox), examples that undermine the deepest foundations of our knowledge (eg. Russell's Paradox, Gödel's Incompleteness Results), and examples that fall somewhere in between. Some paradoxes are based in a natural language only (the Liar Paradox, the White Horse Paradox), some can only be constructed at all with a specific theoretical framework in mind (the Ravens Paradox, the Lottery Paradox).

My account of thought experiments is a generalization of the account Kuhn gives in 'A Function for Thought Experiments' (Kuhn 1977). For Kuhn, a thought experiment can't deliver new information about the world to its experimenter, but it can show tensions concealed within their existing conceptual apparatus. Kuhn claims that thought experiments are fictional scenarios that can reveal the inadequacies of the scientific paradigm of the person who conducts them. I argue that, by the same token, thought experiments can also show that the way the scientist is conceiving of a that scenario *is* adequate to the task at hand. The absence of a failure is a success. On this account, a thought experiment just is any attempt to mentally 'try out' a way of representing a scenario in a realistic use-case in order to show whether or not your current conceptual and representational cognitive machinery can handle the task or not. The conclusion of a thought experiment is a conclusion about the suitability of your representational scheme for this hypothetical task, not a fact about the world. Whether paradoxes have the same epistemic properties as thought

experiments depends on the theory of thought experiments one holds. For instance, on accounts of thought experiments that consider visual imagery or imagined possible worlds to be constitutive of thought experiment, only some paradoxes qualify. A linguistic paradox like Gongsun Long's White Horse paradox requires no mental imagery at all. The paradox is wholly contained in the sentence 'White horse is not a horse' (**白馬非**馬). The Liar Paradox is similar – the only imagined aspect of the paradox is the speaker[65]. So, on an account of thought experiments that requires imaginative mental imagery, neither would qualify. However, both of these paradoxes, after a light argumentative reconstruction, would qualify as thought experiments on John Norton's Argument View of thought experiments since both can be cast into the form of an argument (Norton 2004). Roy Sorensen also notably defends the claim that paradoxes belong to the class of thought experiments. So, whether or not paradoxes count as thought experiments depends on the account of thought experiments to which one subscribes. On my account, visual imagery is present in some but not all thought experiments. If the representation that is to be tested is a linguistic or algebraic representation, then no visual imagery is helpful. Likewise, argument has no more of a special role to play in thought experiments than it does in laboratory experiments. Paradoxes serve the same testing function as thought experiments, and thus should be thought of as such.

If my account is correct and paradoxes are thought experiments (and thus, defined functionally by the goal of testing representations) some of the puzzling aspects of paradoxes become clear. This account demystifies Quine's observation of the disunity of paradoxes from 'Ways of Paradox' (1966). In the essay, Quine points out the vast diversity of the class of canonical paradoxes, which contains everything from merely surprising facts (like that a man born on February 29th would have fewer birthdays than years of age) to deep antinomies such as Russell's paradox.

[65] And even the speaker can be omitted with a sufficiently careful construction. Consider Quine's construction "'Yields a falsehood when appended to its own quotation' yields a falsehood when appended to its own quotation" (Quine 1966)

Quine separates the canonical slate of paradoxes into three kinds: veridical paradoxes, falsidical paradoxes, and antinomies. Veridical paradoxes are constructions that render a surprising but true conclusion (such as the Birthday Paradox); Falsidical paradoxes are paradoxes that, through the presence of a false premise or improper construction somewhere in their set-up, give rise to a false conclusion (though Quine does not cite it, the aforementioned White Horse paradox is an example of this type); and antinomies are paradoxes that generate an impossible or false conclusion from an apparently properly-constructed set-up. Antinomies are typically of the most interest, since if a correctly constructed scenario renders a genuinely impossible result, something must be awry in the construction itself. Not all historically interesting paradoxes are antinomies, but only antinomies necessitate methodological change.

On the account I suggest, in which paradoxes are thought experiments, an antinomy is a kind of test that goes awry: the experimenter constructs a strange or extreme use case for the normal models of a given discipline and determines that the models cannot, in fact, be used to handle that case. This, claims Quine, is why antinomies are so uniquely troubling to their discoverers. They show us that some aspect of our normal ways of going about representing the world must be faulty, and that the only way to comfortably represent this use case is to change some aspect of our normal procedure.

Quine is right to draw out the connection between antinomies and conceptual revisions. He does not, however, connect the other two kinds of paradox to this account. With the view of paradoxes as thought experiments that I advanced above, we can easily incorporate all three varieties. Thought experiments (including paradoxes) are tests of the capacity of a representational system to represent a given scenario. The three kinds of paradox Quine identifies correspond to the three things that can happen following such a test.

Consider any given attempt to solve a particular paradox. This attempt can go one of two ways if the paradox is well-formed: either the experimenter's conceptual scheme can resolve the given situation (sometimes in a surprising manner), or it cannot. In the former case, one derives what Quine would call a veridical paradox, and no more needs to be done. In the latter case, the experimenter derives an antinomy, and some amelioration of the conceptual scheme is required in order to get out of it. It is also possible for a paradox to not be well-formed; in which case the paradox is falsidical. These three possible conclusions are precisely the three possible results of a thought experiment. A thought experiment can be well or poorly constructed; if the latter, it tells the experimenter nothing (falsidical paradox), if the former, it tells the experimenter one of two things: that the representation in question coherently handles the subject matter that has been presented to it (veridical paradox), or that it fails to do so and must be altered (antinomy).

This is the account of paradoxes that I will use to structure the remaining discussion.

4.1.1 Paradoxes in time

In his 1992 PSA comment "Do Thought Experiments have a Life of their own", Ian Hacking argued that thought experiments do not have lives of their own. Unlike laboratory experiments[66], which are made of real materials that can be reconceptualized and reinterpreted by the changing mind of science, thought experiments are forever fixed in the conceptual framework from which they were created (Hacking 1992). Hacking's argument is that, unlike laboratory experiments, which can be rethought, re-interpreted, replicated, and regretted, thought experiments are static slices of their conceptual contexts. After all, a laboratory experiment presents itself to the

[66] Hacking, like many who discuss thought experiments, contrasts them with 'real experiments'. I take this use to be inaccurate and pejorative, so I shall contrast thought experiments with 'laboratory experiments' throughout.

experimenter as bare sense data in need of interpretation, whereas a thought experiment comes to the thinker already interpreted.

I do not intend to dispute the comparative claim – a thought experiment is embedded in the mind of the one who thinks it in a way that is simply not true of a hunk of matter sitting on a laboratory bench. However, a number of philosophers have pointed out that the historical evidence does not strongly support Hacking's claim that thought experiments do not alter over time. There seem to be many historical examples of thought experiments that are transformed and re-imagined in order to fit different contexts by different people. Alisa Bokulich discusses the ways in which the 'Rocket and Thread' thought experiment, first introduced by E. Dewan and M. Beran to draw out a result of Einstein's Special Theory of Relativity, was reinterpreted by John Bell in the framework of Lorenz' Ether Theory (Bokulich 2001). The situation described in the thought experiment and the conclusion the thought experiment derives is the same in both cases, but the two interpretations of the thought experiment demonstrate a coherence in two different theories. Other examples of reinterpreted thought experiments are John Norton's 'TE/Anti-TE pairs' – pairs of thought experiments on apparently the same set-up that render opposite conclusions (Norton 2004).

For Quine, reinterpretability is a crucial feature of paradoxes. Indeed, paradoxes can even shift their taxonomic type alongside shifts in the conceptual background. He notes that, "the falsidical paradoxes of Zeno must have been, in his day, genuine antinomies. We in our latter-day smugness point to a fallacy: the notion that an infinite succession of intervals must add up to an infinite interval. But surely this was part and parcel of the conceptual scheme of Zeno's day… One man's antinomy is another man's falsidical paradox, give or take a couple of thousand years." (Quine 1966) For Quine, there's no question that, say, the Achilles and Tortoise paradox that Aristotle attributes to Zeno is the same paradox that a modern calculus professor uses to prove a point about

infinitesimals to a first-year undergraduate class, despite the difference in outcome. The paradox of the wheel, which we will investigate shortly, makes that transition twice over. First, in the hands of the Pseudo-Aristotelian author of the Mechanics it is an antinomy. Second, in the hands of Galileo, it is a veridical paradox with a very surprising conclusion indeed. Third, in the hands of Mersenne and others, it is a falsidical paradox.

We have three claims on the table here: Quine's claim that paradoxes are reinterpretable, Hacking's claim that thought experiments are not reinterpretable, and my claim that paradoxes are a kind of thought experiment. At most two of these claims can be right. I think the example I will provide shows that Quine and I are right, and Hacking is wrong. Treating paradoxes as thought experiments *explains* their reinterpretability.

The power of this example in particular is that it introduces a crucial caveat on Quine's method of paradox diagnosis. Both the Galilean and the Mersennian solutions to the Paradox of the wheel, which are separate and incompatible, are still current in their respective literatures. Different people believe one or the other not because they have fundamentally different conceptual schemes, but because they have different *goals*. The three solutions locate the source of the paradox in different places because they rely on different understandings of what the paradox should be depicting. They treat the paradox as experiments on different parts of thought. None is right, and none is wrong – they are answering different questions altogether.

Let's now turn to the case:

4.2 The Paradox of the Wheel

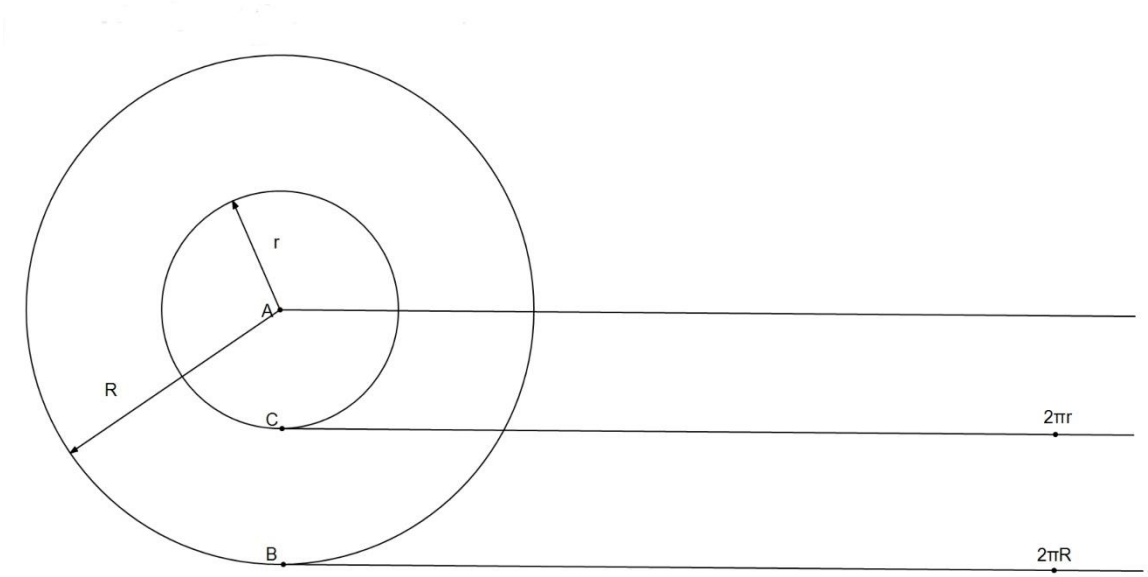

Figure 1: The Psuedo-Aristotelian Wheel

Consider a wheel with a large central hub rigidly connected to its rim by spokes, like a cartwheel. We typically represent a wheel like this this as a large circle (radius R) with a concentric smaller circle inside it (radius r). This wheel is rolling along a rut in a road such that the rim of the wheel is in continuous contact with the bottom of the rut and the hub of the wheel is in continuous contact with the edge of the rut. Both wheels, the inner and outer, are in contact with their respective surfaces at all times. Now, we allow the wheel to turn through one full rotation. How far did it travel? If we measure along the path the rim traced, we find that it traveled one circumference of the larger circle – that is, $2\pi R$. But if we measure along the path the hub of the wheel traced, we find that it too must have traveled one circumference: $2\pi r$. But the two wheels are rigidly attached to

each other and concentric, so whatever arc they traveled must be the same and at the same pace. Therefore, the two wheels must have covered the same distance. This can only be true if $2\pi R = 2\pi r$, which contradicts the premise that $r<R$. We find ourselves in a paradox: the distances traversed by the two wheels must both be equal and not equal.

This is a mathematical paradox of the classic kind, not dissimilar to Zeno's famously unfair races. And like the races, the problem seems to be some sort of disconnect between the mathematical figures used to represent the scenario's constituents and the unproblematic reality of wheels on roads. It is clear to all commentators that something has gone wrong in the description of the wheel, but it is not obvious where the issue lies. In the rest of this paper, I will survey three different solutions to the puzzle – the Pseudo-Aristotelian solution that accompanies the original statement of the paradox in the Aristotelian *Mechanical Questions*, Galileo's matter theory solution from the first day of the Two New Sciences, and the 'Sliding solution' that Galileo puts in the mouth of Sagredo, but which has been taken up by many commentators since, most notably Marin Mersenne. The three solutions are all responses to the same problem. All use the same paradoxical set-up. But all resolve the tension in a different way.

4.3 Reinventing the Wheel

4.3.1 Aristotle's Wheel

The Aristotelian *Mechanical Questions* is structured as a list of question-and-answers, each answer building upon the previous ones. Though it is the opinion of the majority of modern Aristotle scholarship[67] that the author of the *Mechanical Questions* is not Aristotle, its author is working

[67] The originator of this claim in current Aristotle scholarship was W.D. Ross in the early 20th century, but the roots of the dispute over the authenticity of the Mechanics is much older. There remain a few modern Aristotle scholars, such as James Lennox, who believe the Mechanical Questions may be a genuine work of Aristotle (Lennox, personal communication, May 2021)

with close attention to the system of Aristotelian physics and is widely believed to be an unknown member of Aristotle's school[68]. The Rota Aristotelica may not even have been a novel paradox to the author[69], though the *Mechanical Questions* is the earliest known source.

The Mechanical Questions are built around a discussion of the circle, and many of the questions are questions arising from the intersection of linear and circular motion in simple machines. The author claims that circles are a natural source of the paradoxes of motion. Circles intrinsically are made up of contrary properties – motion and stability, concavity and convexity, moving forwards and backwards in the same motion. Circular motion is also not reducible to rectilinear motion, and their interactions are necessarily strained and strange. So, it is only to be expected that a mechanic will have to do careful work to disentangle these potentially paradoxical properties from each other to arrive at useful principles of mechanical motion. The author writes:

> "Now the original cause of all such phenomena [levers and balances] is the circle; and this is natural, for it is in no way strange that something remarkable should result from something more remarkable, and the most remarkable fact is the combination of opposites with each other. The circle is made up of such opposites, for to begin with it is composed both of the moving and of the stationary, which are by nature opposite to each other. (…) This, then, is one peculiarity of the circle, and a second is that it moves simultaneously in opposite directions; for it moves simultaneously forwards and backwards, and the radius which describes it behaves in the same way; for from whatever point it begins, it returns again to

[68] Thomas Nelson Winter identifies the anonymous author with Archytas of Tarentum. However, his evidence is necessarily circumstantial (Winter 2007).

[69] (Drabkin 1950) argues that the wording of the Mechanical Questions suggests that the problem would have been familiar to its audience already. The author of the Mechanical Questions typically opens each problem with a simple 'Διὰ τί', but Problem 24, the Rota Aristotelica, opens with 'Ἀπορειαι διὰ τί' instead, which suggests Problem 24 was a known **ἄπορῖᾶ** (problem or puzzle) within the literature.

the same point; and as it moves continuously the last point again becomes the first in such a way that it is evidently changed from its first position." (847b-848a) (Aristotle 1936)

The author is not claiming that circles themselves are rolling contradictions, but that their properties must be carefully applied so as to massage out the apparent contradictions that emerge from their contrary elements. The paradox of the wheel is a thought experiment that does just this: the author presents a scenario in which the properties of the circle appear to give rise to a contradiction, and with a bit of careful manipulation, show that the contradiction was merely apparent.

Within the Aristotelian paradigm of the *Mechanical Questions*, the relationship between geometry and mechanics is extremely close: mechanics is one of the handful of sciences that Aristotle describes as 'subordinate' to particular branches of mathematics. Harmony is subordinate to arithmetic, optics to geometry, and mechanics to stereometry. The demonstrations of a science that is subordinate to another may, in defiance of usual Aristotelian restrictions, use premises from both the subordinate and superior science. Only demonstrations from the premises of the superior science can reach the reasoned fact (the fact along with its explanation), but demonstrations from the premises of the subordinate science can still reach the fact itself (79a2-15). So, mechanical demonstrations performed in on this Aristotelian model may include premises taken both from geometry and from physics. And indeed, the problems in the *Mechanical Questions* rely heavily on both.

Problem 24, presented in this text without title, seems at first glance to be unrelated to the problems around it. However, all of them, in some way or another, deal with either the paradoxical aspects of circular motion or problems arising from the combination of more than one motion,

often in the context of levers and balances. After constructing the paradox as I presented it above, the author gives his statement of why, exactly, the paradox is such a problem:

> "As, then, nowhere does the greater stop and wait for the less in such a way as to remain stationary for a time at the same point (for in both cases both are moving continuously), and as the smaller does not skip any point, **it is remarkable that in the one case the greater should travel over a path equal to the smaller, and in the other case the smaller equal to the larger**. It is indeed remarkable that as the movement is one all the time, that the same centre should in one case travel a large path and in the other a smaller one. For the same thing travelling at the same speed should always cover an equal path; and moving anything with the same velocity implies travelling over the same distance in both cases." (855b)[70]

So, for the Pseudo-Aristotelian author, there are two problems to consider. First, that it is absurd for the two paths traced to be of the same length without one or the other skipping or sliding (which has been stipulated to be the case). This is just a statement of the paradox, common amongst all accounts. But the second problem he identifies is a distinctively Aristotelian one: if the same moving force is applied to the wheel, it could roll either the circumference of the large or small wheel. If this is the case, there is a crucial ambiguity in the mechanical principles the author is trying to use to solve puzzles – the same force applied to the same wheel can move the wheel either the

[70] Winter's alternative translation of the same passage makes the nature of the puzzle somewhat more obvious: *"It is absurd, with the smaller one [the inner circle] not leaping any point, for the larger [the outer circle] to have gone out an equal extent to the smaller and the smaller and equal extent to the greater. Further, it is marvelous that, with always but one moving force, the center getting moved sometimes rolls out like the large circle, sometimes like the small one. For the thing getting moved at the same speed inherently goes an equal line. And at the same speed it is possible to move it equally either way."* (Winter 2007)

circumference of the smaller or the larger depending on an arbitrary choice of the mechanic describing the motion. One set-up gives rise to two results – the one where the wheel moves a distance of R and the one where the wheel moves a distance r – and the two results cannot be equivalent.

The solution given in the Mechanics is that the two scenarios described above are in fact not the product of the same set-up, only apparently so. The author first notes that in general, when one object moves another, the second shares in most, but not all, of the motion of the first. So, if one body moves according to its own natural motion and in doing so pushes another along, that second body will move in a way unnatural to it but similar to the movement of the first. The distinction between 'moved' and 'mover' is a familiar hallmark of Aristotle's physics, as is the idea that each body has its own distinctive natural motion. The author uses these two tools from his Aristotelian playbook to dissolve the apparent contradiction of the Rota Aristotelica. He explains that the two scenarios identified above – the one in which the wheel rolls the circumference of the larger wheel and the one in which it rolls the circumference of the smaller – are actually not the same scenario at all. The first proceeds from the natural motion of the larger wheel, which moves the smaller wheel unnaturally along with it, and the second proceeds from the movement of the smaller wheel, which likewise brings the larger unnaturally along for the ride. Even though the two circles are concentric and rigidly attached to each other, they are no more the same wheel than if they were unconnected and the one was pushing the other. Their connection, the author points out, is merely an accidental feature of the two wheels, not an essential one. That a strange kind of accidental scenario can force a wheel to roll in an unnatural way is a mere curiosity, not a problem for the applicability of the principles in the rest of the *Mechanical Questions*.

Thus, the paradox is resolved: the contradiction in the wheel set-up was only apparently there because we failed to make the necessary distinction between mover and moved. The problem was that one motion could give rise to two different distances rolled, the solution was that there was never only one motion. For our Pseudo-Aristotelian author, the problem only emerges when we try to look at the problem in a purely geometric context divorced from physics, and the problem dissolves as soon as we add some basic Aristotelian physical principles to its geometric skeleton.

However, the solution that the Author finds rides on the distinctions of Aristotelian physics – between mover and moved, natural and unnatural motion, accidental and essential properties. The statement of the paradox that the author give relies on these concepts, and these concepts are necessary for the resolution. Most modern commentators reject the Aristotelian solution for this reason alone. The Renaissance mathematician Jerome Cardan[71], however, rejected the Aristotelian solution on the grounds that it introduced a physical principle into what he considered to be a purely mathematical problem (Drabkin 1950). Cardan's point is well-taken: the Aristotelian solution is neither wholly a physical nor wholly a mathematical solution. Aristotelian mechanics is a mixed mathematical discipline after all, itself neither wholly mathematical nor physical.

In Quine's taxonomy, this resolution is in the mode of an antinomy, similar to Russell's paradox. The cause of the problem was the ambiguity of the diagram. Since it could not represent the crucial physical facts of the scenario (ie. Which of the wheels was the mover and which was the moved) it could not render an answer to the questions we asked of it without generating a contradiction. The solution is to learn that geometric diagrams only tell part of the story, and that physical principles must be called in to differentiate.

[71] Not coincidentally, Cardan is the first person known to have expressed doubt about the authenticity of the Mechanics (van Leeuwen 2016)

4.3.2 Galileo's Wheel

The first clue that Galileo's treatment of the Paradox of the Wheel will be different is that it is presented very differently within the text in which it occurs, on the first day of the *Two New Sciences*. The first day concerns problems of scale, cohesion, and matter, presented in a loose, free-wheeling dialogue. The conceptual links between the parts are very intricate, and often subtle and hard to discern. In this dialogue, the Rota Aristotelica first occurs in the context of a discussion of the possibility of the vacuum, which progresses into a discussion of paradoxes of the infinite and infinitesimal. Galileo spins several big conceptual plates at once in this section – he not only proposes a significant and controversial metaphysical thesis in avowing the vacuum, but also introduces the suite of mathematical representations he will require to begin to work with this claim later in the book.

Galileo's treatment of the Rota Aristotelica is, in Quine's parlance, a veridical paradox. He fully adopts the paradoxical situation as possible, meaningful, and instructive. On Galileo's treatment, the conclusion of the paradox is highly surprising. Let us now turn to that surprise.

Galileo's solution to the Rota Aristotelica is in the form of a limit proof. First, Galileo invites us to consider an unparadoxical situation, then he shows that the Rota Aristotelica set-up is a limiting case of that situation, and finally he argues that the conclusion to the unproblematic situation is also the proper solution to the problematic one. First, he asks us to consider the situation of a hexagonal 'wheel' constructed in the same manner as the Rota Aristotelica.

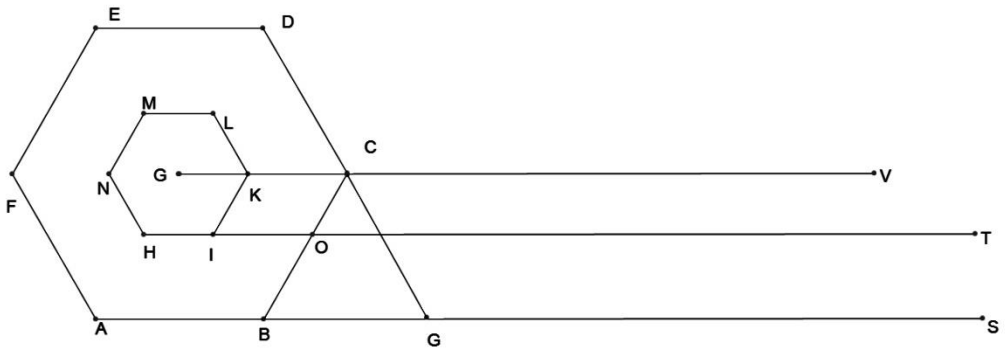

Figure 2: Galileo's Hexagonal Wheel

The 'wheel' can 'roll' along the double tracks AS and HT by pivoting along its corners. First, it pivots around point B, momentarily lifting into the air before coming back down on side BC. Once there, it can pivot again around point C, et cetera. Each time the large hexagon ABCDEF pivots in this way, the small hexagon HIKLMN briefly breaks contact with the track HT before landing again on one of its own faces. Now, imagine that the sides of the hexagons are coated in paint, so we can measure the total length of the line segments traversed by the hexagon in a full rotation. As the hexagonal wheel rolls along, the path laid down in paint by the large wheel will be equal in length to the perimeter of the large hexagon ABCDEF, and the path laid down by the small wheel will be six separated line segments that sum to the perimeter of the small hexagon HIKLMN, with six gaps in the line where the inner wheel lost contact with the upper track. One can also perform the procedure in reverse by pivoting the small hexagon along its track, in which case the large hexagon will break contact with its track and skid back over the line it has already laid down. Again, the total length of the lines laid down by each hexagon is equal to its perimeter, but in the case of the lower hexagon, some of the lines laid down overlap each other.

The situation with the hexagon is unparadoxical and easily imagined. Next, Galileo asks us to imagine the same procedure, but with a pair of chiliagons for the wheel instead of hexagons. The chiliagons behave just like the hexagons did, with the large chiliagon pivoting on the lower track and the small one briefly losing contact on each of the large chiliagon's thousand pivots. At this point, Galileo suggests, the hops made by the chiliagon wheels will be so small and frequent as to be imperceptible, though they are still of finite size. Yet, it is still the case that the line laid down by each chiliagon is equal to its perimeter, with the caveat that the line laid down by the smaller chiliagon is full of imperceptible gaps. And, indeed, though in the case of the hexagon the lines AS, HT, and GV, which represent the tracks traversed in a full rotation by the larger hexagon, smaller hexagon plus gaps, and center axis plus gaps respectively, were of slightly different sizes, in the case of a chiliagon the individual sides are so small that the three lines would be virtually identical in length.

So, the more sides in the polygonal wheel, the closer the total distance traversed, and the smaller the hops. And what is a circle but a polygon with infinitely many sides?

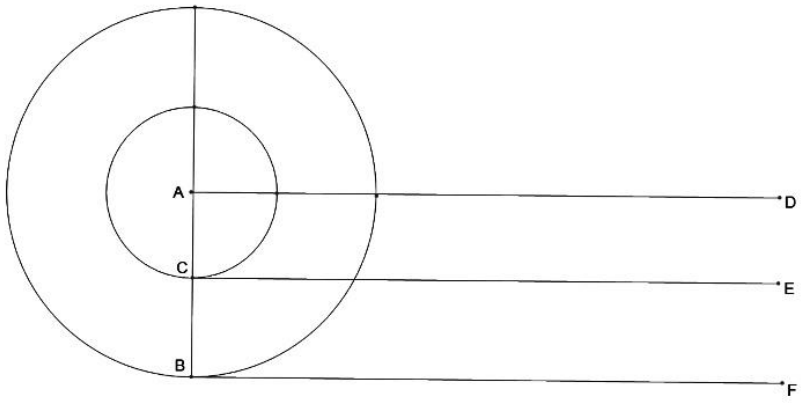

Figure 3: Galileo's Circular Wheel

Thus, we arrive at Galileo's depiction of the paradox of the wheel. The larger wheel is pivoting on all its points along the track BF like the polygons did upon their corners, and the inner wheel is making innumerable tiny hops off the line CE. Both the pivots and the gaps are now infinitesimally small. Yet, as Galileo argues, the result from the hexagons holds just the same: the large wheel lays down its own circumference, and the small wheel lays down its own circumference interspersed with infinitesimal gaps that, taken together, add up to the difference between the circumferences of the two wheels. So, claims Galileo, the paradox is resolved: the large wheel traverses $2\pi R$ and the small one traverses $2\pi r$, with no slipping, despite the fact that the two wheels are rigidly connected and turn at the same rate.

The consequence of Galileo's resolution of the paradox, however, is that we now must countenance that two lines that appear to be the same length on paper may not be the same length at all. Galileo leans into the consequence completely. He says, in the triumphant voice of Salviati:

> "But if we consider the line resolved into an infinite number of infinitely small and indivisible parts, we shall be able to conceive the extended indefinitely by the interposition, not of a finite, but of an infinite number of infinitely small indivisible empty spaces.
>
> **Now this which has been said concerning simple lines must be understood to also hold in the case of surfaces and solid bodies**, it being assumed that they are also made up of an infinite, not a finite, number of atoms." (Galilei 1914) (First day, 72. Trans. Crew and Da Salvio)

Galileo spends the next thirty pages defending the extraordinary claim he has just made. The claim that all matter is divisible into infinitely many infinitesimal atoms requires a great deal of conceptual finesse, and Simplicio and Sagredo put Salviati through his paces to preserve the conceptual coherence of the view. Eventually, Galileo comes to a fleshed-out method for discussing

the infinite and the indivisible – infinitely large and infinitely small things cannot be compared to other infinites in size and number, since all infinite quantities can be set in a one-to-one correspondence with each other, just as roots and their squares can. Galileo's discussion here seems to presage Cantor, and that Cantorian flavour has seeped into many modern analyses of Galileo's solution to the Rota.

The discussion of how to handle and work with infinite and infinitesimal quantities also works its way into Galileo's own discussion of the Rota. A seldom-cited fact about Galileo's solution to the paradox is that it occurs twice in the first day of the Two New Sciences. The first occurrence is more thorough and presents 'the Galilean Solution' most clearly. Until now, the first presentation is the one I have explicated. However, the second occurrence of the solution to the Rota Aristotelica is the one that gives us a clue about Galileo's broader aims. Let us now discuss them.

The conclusion that Galileo gives for the first presentation of the Rota Aristotelica was the existence of infinitesimal atoms and vacua; the conclusion for the second is the related but distinct claim that any finite quantity can be divided into infinitely many infinitessimal parts and that any two such finite quantities can be set in correspondence with each other. In the thirty-odd pages between the two presentations of the Rota, the concern has changed from being largely metaphysical to largely methodological.

As Olympia Nicodemi points out, the methodological upshot of being able to put two quantities that are subdivided into infinitesimal parts into a one-to-one correspondence of their infinite parts to one another is a result that Galileo will require later in the Two New Sciences in his proofs of the law of fall. Both the disproof of the speed-distance law of fall and the demonstration of the true time-squared law of fall require Galileo to divide a continuous quantity of time into infinitesimal chunks and to compare them at each step. Nicodemi argues that the crucial upshot of

the Rota Aristotelica proof is that it provides Galileo with a justification for the infinitesimal methods he will use in the later days (Nicodemi 2014).

Galileo is quite explicit that his second invocation of the Rota Aristotelica is for methodological ends. After a back-and-forth between Simplico and Salviati over whether the division of continuous quantities into infinitesimal parts is actual or merely potential, Simplicio throws down the gauntlet – after all, if dividing a line into infinitely many segments is impossible in practice, how can it possibly be actual?

> SIMP: I cannot help admiring your discussion; but I fear that this parallelism between the points and the finite parts contained in a line will not prove satisfactory, and that you will not find it so easy to divide a given line into an infinite number of points as the philosophers do as to cut it into ten fathoms or forty cubits; not only so, but such a division is quite impossible to realize in practice, so that this will be one of those potentialities that cannot be reduced to actuality.
>
> SALV: The fact that something can be done only with effort or diligence or with great expenditure of time does not render it impossible; for I think that you yourself could not easily divide a line into a thousand parts, and much less if the number of parts were 937 or any other large prime number. But if I were to accomplish this division which you deem impossible as readily as another person would divide the line into forty parts would you be more willing, in our discussion, to concede the possibility of such a division? (Galilei 1914)

The method of dividing a line into infinitely many parts that Salviati is proposing is the Rota Aristotelica again, a procedure that separates the constituent part of a line into infinitely many infinitesimal parts by stamping them down along a track interspersed with infinitely many corresponding infinitesimal vacua. This in-principle division of the line is highly salient to Galileo's

broader goals. As Nicodemi notes, Galileo will resolve continuous lines into sequences of infinitesimals on many other occasions throughout the Two New Sciences.

In one paradox, Galileo has advanced both his theory of the continuum and the attendant method of resolving continuous quantities into their infinitesimal components. In Quine's terminology, Galileo's interpretation of the Rota Aristotelica is a veridical paradox, and an uncommonly fruitful one at that. Galileo's solution, under a newer Cantorian gloss, is still commonly given as the solution to the Rota Aristotelica in modern treatments[72].

However, not all were satisfied with Galileo's solution to the Rota Aristotelica, including Galileo's own fictional interlocutors. Sagredo and Simplicio both express their doubts about the solution that Salviati suggests. Sagredo worries near the start of the discussion that, contrary to the premise of the paradox, "the points on the circumference of the small circle, carried along by the motion of the larger circle, would slide over some small parts of the line" (70). Simplicio's worry is more general – after Salviati has given the complete proof, he throws up his rhetorical hands and says that "the arguments and demonstrations which you [Salviati] have advanced are mathematical, abstract, and far removed from concrete matter; and I do not believe that when applied to the physical and natural world these laws will hold." (96). Galileo's eye for objections to his work is here as keen as ever: both these contentions are crucial in animating the third and most recent solution to the paradox that we shall consider. Let us turn to it next.

[72] Wolfram Mathworld, a reference resource commonly used by mathematics and physics students, gives only the Galilean solution to the Rota Aristotlica without mention of any other solutions. (Eric W. Weisstein n.d.)

4.3.3 Mersenne's Wheel

> "Real experiment resolves Problem 24 of Mechanical Problems completely—provided we agree on what the problem is!" -Richard TW Arthur (Arthur 2012)

It might be suggested that wheels and roads are perfectly sensible physical objects. There is no special impossible element contained within the set-up of this thought experiment, just wheels and roads. The wheel is idealized, with its perfectly circular wheels meeting their perfectly flat tracks at a single point each, and perfectly rigid connections between the two wheels. These idealizations are well-within the bailiwick of normal physics, though – perfectly familiar to any advanced high schooler. There is no reason those idealizations should preclude the methods that physicists normally use to solve problems of this kind. We want to find out how to resolve the discrepancy between the lengths of the paths traced by the two circumferences as they turn? All we need to do is get some wheels and a track and check. The author of the Mechanics did not have access to the high-speed cameras and precision-engineered wheels that we have now, so it's no fault of his that he didn't try it. But we live in an age where the solutions to all wheel-based problems should be within our grasp. As Marin Mersenne argues in his influential presentation of the paradox:

> "But we must admit that the negligence of men is strange, that it is so often mistaken for not wanting to have the least experience of the world and works itself in vain searching for reasons for something that has no point, as happens in this case, because the small circle never moves the large one except when the many parts of the large circle do not touch the same part of the plane, such that each part is touched by one hundred different parts of the large circle when it is one hundred times as large as the other. And when the little one is moved by the large one, the same parts of the small touch a hundred parts of the large, **as**

experience will show to all those who do the experiment in a big enough volume[73]."

(Mersenne, Trans. Jennifer Whyte and John Buchanan)

Mersenne points out that the impossibility of the set-up becomes clear to anyone who cares to check. If you build a physical double wheel as described in the paradox, give it a spin, and observe it sufficiently carefully, you will immediately see that Sagredo's worry has been vindicated. The smaller wheel does not evenly rotate on its surface, but instead slides and skips. The paradox seems, then, to be a moot point: the idealized scenario described in the set-up of the thought experiment does not obtain, so there is no problem with the impossible consequences of supposing it does. Real-world cartwheels and roads are safe from any contradictions in their movement – the whole paradox, on this view, was merely an artefact of the idealizations used to depict it.

This response has a history nearly as long as the history of the paradox itself. The author of the Mechanical Questions explicitly rules it out in the set-up of the paradox. Galileo puts the response in the mouth of Sagredo and then immediately argues against it at length. However, many authors (such as the afore-quoted Mersenne) took it up after Galileo, following Mersenne's translation of the Two New Sciences. These accounts continue to the present day. A canonical (and more thorough) presentation of the sliding solution forms the entry for 'Rota Aristotelica' in Charles Hutton's *Mathematical and Physical Dictionary* of 1795, which he attributes to the French natural philosopher Jaque-Jean d'Ourtous de Mairan:

[73] "*Or il faut advouer que la negligence des hommes est etrange, qui se trompent si souvent pour ne vouloir pas faire la moindre experience du monde & qui se travaillent en vain à la recherche des raisons d'une chose qui n'est point, comme il arrive en celle cy, car le petit cercle ne meut iamais le grad que plusieurs parties du grand ne touchent une mesme partie du plan, dont chaque partie est touchee par cent parties differentes du grand cercle quand il est cent fois plus grand que l'autre. Et lors que le petit est meu par le grand, une mesme partie du petit, touche cent parties du grand, comme l'experience fera voir à tous ceux qui la feront en assez grand volume*." -Marin Mersenne, *Les Mechaniques de Galilee mathematicien, traduites d'Italien per Pere Mersenne*, quoted in (Drabkin 1950). Original translation.

> Rota Aristotelica, or Aristotle's Wheel, denotes a celebrated problem in mechanics, concerning the motion or rotation of a wheel about its axis; so called because first noticed by Aristotle[74].
>
> (…)
>
> After the fruitless attempts of so many great men, M. Dortous de Meyran, a French gentleman, had the good fortune to hit upon a solution, which he sent to the Academy of Sciences; where being examined by Mess. de Louville and Soulmon, appointed for that purpose, they made their report that it was satisfactory. The solution is to this effect:
>
> The wheel of a coach is only acted on, or drawn in a right line; its rotation or circular motion arises purely from the resistance of the ground upon which it is applied. Now this resistance is equal to the force which draws the wheel in the right line, inasmuch as it defeats that direction; of consequence the causes of the two motions, the one right and the other circular, are equal. And hence the wheel describes a right line on the ground equal to its circumference.
>
> As for the nave of the wheel, the case is otherwise. It is drawn in a right line by the same force as the wheel; but it only turns round because the wheel does so, and can only turn in the same time with it. Hence it follows, that its circular velocity is less than that of the wheel, in the ratio of the two circumferences; and therefore its circular motion is less than the rectilinear one. Since then it necessarily describes a right line equal to that of the wheel**, it can only do it partly by sliding, and partly by revolving**, the sliding part being more or less as the nave itself is smaller or larger. See Cycloid. (Hutton 1795)

Though Hutton's presentation is not original, it is instructive. For Hutton, the solutions of the Aristotelian author and Galileo (and of Taquet, omitted) are not merely false solutions but no solutions at all. Aristotle merely restates the difficulty, and all of Galileo's mathematical finesse is

[74] Aristotle was still universally believed to be the author of the mechanics until the early 20th Century. See earlier notes for discussion of the authorship of the Mechanics.

irrelevant to the physics at hand. Hutton's framing is clear: this is a problem of physics, to be solved with physics alone. Anything else misses the point.

There is, however, a caveat to this Sliding Solution. As Arthur notes in his discussion of whether or not thought experiments taken as a general class can be resolved by experiment (in which he discusses the Rota Aristotelica as an example), it is a solution to a different kind of problem to the one raised in the previous two solutions. The nature of the object at issue has changed, from idealized wheels on paper to real wheels in a laboratory (Arthur 2012). This is sufficient to assuage any fears we may have had for the movement of real wheels, but it will not solve the paradox on paper. Nor is it sufficient, as a solution, to explain what the proper relation between the depiction of the wheel on paper and the wheel on the ground should be. One cannot derive the Sliding Solution from the geometry of circles and lines, one must either posit it to explain away the paradox or observe it empirically. It is no more a solution to the mathematical paradox than noticing how tortoises and invulnerable Greek demigods actually move is a solution to Zeno's paradoxes of motion. Insofar as the Sliding Solution is an answer to anything, it is an answer to the question of why the paradox does not bother us, not an answer to the question raised by the paradox itself. It shows us that the problems in our representation of the rotation of wheels and as circles pivoting around particular points do not carry over into the behaviour of real rotating wheels – that something in the set-up is wrong. It does not, however, provide an alternative way of representing wheels that does not generate this problem. The Mersenne solution depicts the Rota Aristotelica as, on Quine's taxonomy, a falsidical paradox. It denies that the situation depicted in the paradox, a situation in which there is no slipping, is well-formed. Thus, like the classic falsidical paradoxes, the Rota Aristotelica teaches us nothing.

Like the Galilean solution, Mersenne's solution is still commonly given as the answer to the Rota Aristotelica in modern sources (for an example, see (Bunch 1997)). This fact is not surprising: the Sliding solution is the correct answer in a particular type of practice – namely, the practice of applying Newtonian dynamical principles to slightly idealized physical systems. Any reader with an undergraduate physics background likely attempted to begin to figure out the relevant forces on the wheels as soon as the problem was presented. And the falsidical solution – that in this kind of context the paradox has an incoherent set-up – is exactly what one derives when one does so.

4.4 Conclusion

This chapter has focused on two large themes.

First: Thought Experiments are methods for determining the ability of representational scheme to be used for specific purposes. Paradoxes are a kind of thought experiment, so they have the same function. Quine's tripartite division of paradoxes captures the three possible results of a thought experiment: failure of the thought experiment (falsidical paradox), successfully proving the representational scheme can be used for the given purpose (veridical paradox), and proving that the representational scheme contains some deep problem that prevents it from being used for the given purpose (antinomies). Characterizing Quine's trifold taxonomy as three outcomes of the same process reunifies the domain of paradoxes.

Second: Paradoxes and other thought experiments can be reinterpreted in different contexts for precisely this reason. Because thought experiments test representational schemes by attempting to use a particular representation in a particular operational context, changes in that context result in different outcomes to the same thought experiment. These contextual changes are not limited to

Kuhnian paradigm changes in science but can also occur between different disciplines and different kinds of scientific goal.

The example of the Rota Aristotelica demonstrates these two themes. The long and winding history of the paradox features three different problem contexts, and three different ways of escaping them. For the Aristotelian author of the Mechanics, the incoherence occurred in the ambiguous use of a diagram, and was solved by noting that the diagram actually represented two different physical scenarios. For Galileo, the incoherence came from an incomplete understanding of the nature of the continuum, which he solved by introducing his distinctive brand of atomism and new methods of mathematically reckoning the indivisible. Mersenne and many subsequent authors saw the incoherence as an unphysical description of an impossible wheel, and resolved the tension by observing its incompatibility with experience and leaving it at that. Each response is a reasonable response to its own circumstance of use, but no answer will suffice to answer another author's question. There is no contradiction here, no tension. Just two wheels rolling on to two different destinations.

Conclusion

One notable feature of the account of thought experiments that I have elaborated and defended within this dissertation is that the function that I identify with thought experiment is very common. If I'm right, thought experiment is a near-ubiquitous part of how thinkers change their minds. We are all constantly grinding the lenses we use to see the world, and flipping back and forth between them like an optician. Were we not, we wouldn't be able to see at all.

At the beginning of this dissertation, I characterized scientific practice as a constant churn of novelty. The guarantee that our descriptive strategies work is constantly undermined by the shifting meanings of the concepts from which they are composed. Science requires a mechanism by which these strategies can be checked for descriptive adequacy – an analogue to a check on logical validity for non-propositional systems. These non-empirical checks are vital to the continuing function of science. I call these checks 'thought experiments', and I think the majority of the procedures currently called thought experiments can be fruitfully analyzed as non-empirical stress-tests of this kind. By insisting on analyzing thought experiments as serving the same function as laboratory experiments (that is, the production of evidence for or against a claim about the world) previous accounts of thought experiments have mischaracterized and obscured the role they truly play in science. Throughout the remainder of this dissertation, I demonstrated the ability of this account to make sense of many cases of thought experiments in the history of science, including a few novel ones. These cases are not alone – I hope that the account I have given opens the door to greater understanding of science at its very strangest.

The End.

Bibliography

Aristotle. 1936. Minor Works: On Colours. On Things Heard. Physiognomics. On Plants. On Marvellous Things Heard. Mechanical Problems. On Indivisible Lines. The Situations and Names of Winds. On Melissus, Xenophanes, Gorgias. Translated by W. S. Hett. Vol. 307. Loeb Classic Library. Cambridge, Massachusetts: Harvard University Press. https://www.loebclassics.com/view/LCL307/1936/volume.xml.

Arthur, Richard T. W. 2012. "Can Thought Experiments Be Resolved by Experiment?: The Case of Aristotle's Wheel." In Thought Experiments in Science, Philosophy, and the Arts. Routledge.

Baritompa, Bill, Rainer Löwen, Burkard Polster, and Marty Ross. 2018. "Mathematical Table Turning Revisited." arXiv. http://arxiv.org/abs/math/0511490.

Batterman, Robert W. 2013. "The Tyranny of Scales." In The Oxford Handbook of Philosophy of Physics, edited by Robert W. Batterman, 255–86. Oxford University Press.

Blackburn, Simon. 2008. "Open Texture." In The Oxford Dictionary of Philosophy. Oxford University Press. https://www.oxfordreference.com/view/10.1093/acref/9780199541430.001.0001/acref-9780199541430-e-2252.

Boethius, Anicius Manlius Severinus. 2009. The Consolation of Philosophy. Harvard University Press.

Bokulich, Alisa. 2001. "Rethinking Thought Experiments." Perspectives on Science 9 (3): 285–307. https://doi.org/10.1162/10636140160176152.

Boltzmann, Ludwig. 1895. On Certain Questions of the Theory of Gases. Springer Nature. http://archive.org/details/paper-doi-10_1038_051413b0.

———. 2015. "'On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium' Sitzungberichte Der Kaiserlichen Akademie Der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, Pp 373-435 (Wien. Ber. 1877, 76:373-435). Reprinted in Wiss. Abhandlungen, Vol. II, Reprint 42, p. 164-223, Barth, Leipzig, 1909." Translated by Kim Sharp and Franz Matschinsky. Entropy 17 (4): 1971–2009. https://doi.org/10.3390/e17041971.

———. 2021. "Boltzmann's Philosophy Notes for Three Lectures (Fall 1903)." LUDWIG BOLTZMANN, 13.

Boyle, Robert. 1999. The Works of Robert Boyle: "The Origin of Forms and Qualities" and Other Publications of 1665-7. Pickering & Chatto.

Brown, Harvey R., Wayne Myrvold, and Jos Uffink. 2009. "Boltzmann's H-Theorem, Its Discontents, and the Birth of Statistical Mechanics." Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 40 (2): 174–91. https://doi.org/10.1016/j.shpsb.2009.03.003.

Brown, James Robert. 2010. The Laboratory of the Mind: Thought Experiments in the Natural Sciences. https://www.routledge.com/The-Laboratory-of-the-Mind-Thought-Experiments-in-the-Natural-Sciences/Brown/p/book/9780415996532.

Bunch, Bryan. 1997. Mathematical Fallacies and Paradoxes. Courier Corporation.

Callender, Craig, and Jonathan Cohen. 2005. "There Is No Special Problem About Scientific Representation." Theoria: Revista de Teoría, Historia y Fundamentos de La Ciencia 21 (1): 67–85.

Cantini, Andrea, and Riccardo Bruni. 2021. "Paradoxes and Contemporary Logic." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/paradoxes-contemporary-logic/.

Cartwright, Nancy. 1983. How the Laws of Physics Lie. New York: Oxford University Press.

Cercignani, Carlo. 1998. Ludwig Boltzmann: The Man Who Trusted Atoms. Oxford ; New York: Oxford University Press.

Clausius, Rudolph. 1851. "On the Moving Force of Heat, and the Laws Regarding the Nature of Heat Itself Which Are Deducible Therefrom." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 4, 2.

Collins, H. M. 1991. "The Meaning of Replication and the Science of Economics." History of Political Economy 23 (1): 123–42. https://doi.org/10.1215/00182702-23-1-123.

De Regt, Henk. 1999. "Ludwig Boltzmann's 'Bildtheorie' and Scientific Understanding," 23.

Dennett, Daniel C. 2013. Intuition Pumps And Other Tools for Thinking. W. W. Norton & Company.

Drabkin, Israel E. 1950. "Aristotle's Wheel: Notes on the History of a Paradox." Osiris 9 (January): 162–98. https://doi.org/10.1086/368528.

Ehrenfest, Paul, and Tatiana Ehrenfest. 2014. The Conceptual Foundations of the Statistical Approach in Mechanics. Courier Corporation.

Elgin, Catherine Z. 2014. "Fiction as Thought Experiment." Perspectives on Science 22 (2): 221–41. https://doi.org/10.1162/posc_a_00128.

Eric W. Weisstein. n.d. “Aristotle’s Wheel Paradox.” Text. Mathworld - A Wolfram Math Resource. Wolfram Research, Inc. Accessed May 12, 2022. https://mathworld.wolfram.com/.

Franklin, Allan, and Slobodan Perovic. 2023. “Experiment in Physics.” In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta and Uri Nodelman, Summer 2023. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/physics-experiment/.

Frigg, Roman, and James Nguyen. 2021. “Scientific Representation.” In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/scientific-representation/.

Galilei, Galileo. 1914. Dialogues Concerning Two New Sciences. Macmillan.

———. 1962. Dialogue Concerning the Two Chief World Systems, Ptolemaic and Copernican, Second Revised Edition. Translated by Stillman Drake. 2nd ed.

Gendler, Tamar Szabó. 1998. “Galileo and the Indispensability of Scientific Thought Experiment.” The British Journal for the Philosophy of Science 49 (3): 397–424.

Gendler, Tamar Szabó. 2004. “Thought Experiments Rethought—and Reperceived.” Philosophy of Science 71 (5): 1152–63. https://doi.org/10.1086/425239.

Giere, Ronald N. 2010. Scientific Perspectivism. Chicago, IL: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/S/bo4094708.html.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, and Lorenz Kruger. 1990. The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge University Press.

Gödel, Kurt. 1949. "An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation." Reviews of Modern Physics 21 (3): 447–50. https://doi.org/10.1103/RevModPhys.21.447.

Goodman, Nelson. 1968. Languages of Art: An Approach to a Theory of Symbols. Bobbs-Merrill.

Hacking, Ian. 1975. The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference. Cambridge University Press.

———. 1992. "Do Thought Experiments Have a Life of Their Own? Comments on James Brown, Nancy Nersessian and David Gooding." PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992 (2): 302–8. https://doi.org/10.1086/psaprocbienmeetp.1992.2.192844.

Hesse, Mary B. 1966. Models and Analogies in Science. Ind.

Hutton, Charles. 1795. A Mathematical and Philosophical Dictionary: Containing an Explanation of the Terms, and an Account of the Several Subjects, Comprized Under the Heads Mathematics, Astronomy, and Philosophy Both Natural and Experimental: With an Historical Account of the Rise, Progress, and Present State of These Sciences: Also Memoirs of the Lives and Writings of the Most Eminent Authors, Both Ancient and Modern, Who by Their Discoveries Or Improvements Have Contributed to the Advancement of Them ... With Many Cuts and Copper-Plates. J. Davis.

Kuhn, Thomas S. 1977. The Essential Tension: Selected Studies in Scientific Tradition and Change. Chicago: University of Chicago Press.

———. 2012. The Structure of Scientific Revolutions: 50th Anniversary Edition. University of Chicago Press.

Lakatos, Imre. 1978. The Methodology of Scientific Research Programmes: Philosophical Papers. Edited by John Worrall and Gregory Currie. Vol. 1. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621123.

Lakatos, Imre, John Worrall, and Elie Zahar, eds. 1976. Proofs and Refutations: The Logic of Mathematical Discovery. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139171472.

Leeuwen, Joyce van. 2016. The Aristotelian Mechanics: Text and Diagrams. Springer Verlag.

Lem, Stanislaw. 2002. The Cyberiad. Houghton Mifflin Harcourt.

Machery, Edouard. 2004. "Semantics, Cross-Cultural Style." Cognition 92 (3): B1–12. https://doi.org/10.1016/j.cognition.2003.10.003.

———. 2011. "Thought Experiments and Philosophical Knowledge." Metaphilosophy 42 (3): 191–214. https://doi.org/10.1111/j.1467-9973.2011.01700.x.

Martin, Andre. 2007. "On the Stability of Four Legged Tables." Physics Letters A 360 (4–5): 495–500. https://doi.org/10.1016/j.physleta.2006.08.053.

Miščević, Nenad. 1992. "Mental Models and Thought Experiments." International Studies in the Philosophy of Science 6 (3): 215–26. https://doi.org/10.1080/02698599208573432.

Mitchell, Sandra D. 2002. "Integrative Pluralism." Biology and Philosophy 17: 55–70.

———. 2009. Unsimple Truths: Science, Complexity, and Policy. University of Chicago Press.

Morgan, Mary S., and Margaret Morrison, eds. 1999. Models as Mediators: Perspectives on Natural and Social Science. Ideas in Context. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511660108.

Nersessian, Nancy J. 1992. "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling." PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992 (2): 291–301. https://doi.org/10.1086/psaprocbienmeetp.1992.2.192843.

———. 2008. Creating Scientific Concepts. Cambridge, UNITED STATES: MIT Press. http://ebookcentral.proquest.com/lib/pitt-ebooks/detail.action?docID=3338942.

Nicodemi, Olympia. 2014. "Galileo and Aristotle's Wheel." Journal of Humanistic Mathematics 4 (1): 2–15. https://doi.org/10.5642/jhummath.201401.03.

Norton, John D. 1991. "Thought Experiments in Einstein's Work." In , edited by Tamara Horowitz and Gerald J. Massey. Savage, MD: Rowman & Littlefield. http://d-scholarship.pitt.edu/12664/.

———. 2002. "Why Thought Experiments Do Not Transcend Empiricism." In Contemporary Debates in the Philosophy of Science, edited by Christopher Hitchcock, 44–66. Blackwell.

Norton, John D. 2004. "Why Thought Experiments Do Not Transcend Empiricism," January.

Norton, John D. 2018a. "The Worst Thought Experiment." In The Routledge Companion to Thought Experiments, edited by Michael T. Stuart, Yiftach J. H. Fehige, and James Robert Brown. Routledge.

———. 2018b. "How to Build an Infinite Lottery Machine." European Journal for Philosophy of Science 8 (1): 71–95. https://doi.org/10.1007/s13194-017-0174-4.

———. 2020. "How NOT to Build an Infinite Lottery Machine." Studies in History and Philosophy of Science Part A 82 (August): 1–8. https://doi.org/10.1016/j.shpsa.2019.07.001.

———. 2022. "How to Make Possibility Safe for Empiricists." In Rethinking the Concept of Law of Nature, edited by Yemima Ben-Menahem, 129–59. Jerusalem Studies in Philosophy and History of Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-96775-8_5.

———. forthcoming. "Chance Combinatorics: The Theory That History Forgot."

Polya, George. 1954a. Mathematics and Plausible Reasoning, Volume 1. https://press.princeton.edu/books/paperback/9780691025094/mathematics-and-plausible-reasoning-volume-1.

———. 1954b. Mathematics and Plausible Reasoning, Volume 2. https://press.princeton.edu/books/paperback/9780691025100/mathematics-and-plausible-reasoning-volume-2.

Quine, W. V. O. 1966. The Ways of Paradox. New York: Random.

Roux, Sophie. 2011. "The Emergence of the Notion of Thought Experiments." In , 1. Brill. https://shs.hal.science/halshs-00807058.

Schmitt, Richard Henry. 2011. "Models, Their Application, and Scientific Anticipation: Ludwig Boltzmann's Work as Tacit Knowing." Bulletin of Science, Technology & Society 31 (3): 200–205. https://doi.org/10.1177/0270467611406517.

Shapiro, Stewart, and Craige Roberts. 2021. "Open Texture and Mathematics." Notre Dame Journal of Formal Logic 62 (1). https://doi.org/10.1215/00294527-2021-0007.

Sorensen, Roy A. 1992. Thought Experiments. New York: Oxford University Press.

Stuart, Michael T. 2020. "The Material Theory of Induction and the Epistemology of Thought Experiments." Studies in History and Philosophy of Science Part A 83 (October): 17–27. https://doi.org/10.1016/j.shpsa.2020.03.005.

———. 2023. "Thought Experiments in Chemistry," June. https://doi.org/10.17605/OSF.IO/Y7VXR.

———. forthcoming. "Scientists Are Epistemic Consequentialists about Imagination." Philosophy of Science. https://philarchive.org/rec/STUSAE-3.

Suárez, Mauricio. 2003. "Scientific Representation: Against Similarity and Isomorphism." International Studies in the Philosophy of Science 17 (3): 225–44. https://doi.org/10.1080/0269859032000169442.

Suppes, Patrick. 1962. "Models of Data." In Logic, Methodology and Philosophy of Science Proceedings of the 1960 International Congress, edited by Ernest Nagel, Patrick Suppes, and Alfred Tarski.

Tanaka, Setsuko. 1999. "Boltzmann on Mathematics." Synthese 119 (1/2): 203–32.

Tanswell, Fenner, Colin Rittberg, and Brendan Larvor. 2022. "Lakatos And Kneebone: At The Roots Of A Dialectical Philosophy Of Mathematics." In . London School of Economics.

Thagard, Paul. 2014. "Thought Experiments Considered Harmful." Perspectives on Science 22 (2): 288–305. https://doi.org/10.1162/POSC_a_00131.

Tolkien, J. R. R. 1966. The Tolkien Reader. New York: Ballantine Books.

Uffink, Jos. 2007. "COMPENDIUM OF THE FOUNDATIONS OF CLASSICAL STATISTICAL PHYSICS." In Philosophy of Physics, 923–1074. Elsevier. https://doi.org/10.1016/B978-044451560-5/50012-9.

———. 2014. "Boltzmann's Work in Statistical Physics." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2022. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/statphys-Boltzmann/.

Waismann, Friedrich. 1947. "Verifiability." Journal of Symbolic Logic 12 (3): 101–101. https://doi.org/10.2307/2267243.

Whyte, Jennifer. 2021. "The Roots of the Silver Tree: Boyle, Alchemy, and Teleology." Studies in History and Philosophy of Science 85 (February): 185–91. https://doi.org/10.1016/j.shpsa.2020.10.007.

Wilkes, Kathleen V. 1988. Real People: Personal Identity Without Thought Experiments. Oxford, GB: Oxford University Press.

Wilson, Mark. 2006. Wandering Significance: An Essay on Conceptual Behaviour: An Essay on Conceptual Behaviour. OUP Oxford.

———. 2017. Physics Avoidance: Essays in Conceptual Strategy. Oxford University Press.

Wimsatt, William C. 2007. Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality. Harvard University Press.

Winter, Thomas Nelson. 2007. "The Mechanical Problems in the Corpus of Aristotle." Faculty Publications, Classics and Religious Studies Department. 68.