Faculty of Arts
University of Helsinki

THE EPISTEMIC ROLE OF PREDICTION IN SCIENCE

Pekka Syrjänen

DOCTORAL DISSERTATION

To be presented for public discussion with the permission of the Faculty of Arts of the University of Helsinki, in Auditorium 116, Unioninkatu 35, on the 16th of December 2022 at 13 o'clock.

Helsinki 2022

ISBN 978-951-51-8742-0 (pbk.)
ISBN 978-951-51-8743-7 (PDF)

Unigrafia
Helsinki 2022

ABSTRACT

This study investigates the epistemic role and value of prediction in science: do predictions, and what kind of predictions, justify belief in scientific theories? In the philosophy of science, the epistemic role of prediction has been understood mainly in terms of novel prediction. Scientific theories are considered to gain special epistemic support if they predict novel empirical results that were not used in their construction. This view has played a central role in two debates in the field: prediction vs. accommodation and scientific realism vs. anti-realism. In the former debate, predictivists seek to explain why novel predictions confirm the theory more strongly than accommodations, i.e. empirical results that were used in the construction of the theory. In the latter debate, predictivists argue that novel prediction is one of the most important criteria of empirical success that justifies realist commitment to scientific theories.

Novelty-based predictivism has been defended through two strategies. One strategy is negative: it argues that novel prediction should be preferred because accommodation is associated with negative epistemic consequences. Another strategy is positive: it holds that only novel predictive success calls for a realist explanation. This study evaluates and rejects both strategies. After other epistemic factors and the problems of novel prediction itself are taken into account, there are no general or substantial epistemic advantages to novel prediction due to the problems of accommodation. After the competitive context of scientific practice where novel predictions are pursued is taken into account, there is no need for a realist explanation in cases of novel success. In accordance, novelty-based predictivism is rejected as a viable approach to the epistemic role and value of prediction in science in this study.

In place of novelty-based predictivism, the study advocates another, logical approach to scientific prediction. It is argued that what matters for scientific confirmation is the predictivity or predictive performance of the theory itself. A new predictive virtues approach is introduced, which argues that the predictivity of a scientific theory is a more multifaceted property than has been previously recognized in the philosophical literature. Four predictive virtues are introduced which probe in different ways the predictivity of a scientific theory: veridicality, specificity, scope, and counterfactual depth. The new approach is evaluated against problems and puzzles that face predictivist theories of scientific confirmation. It is argued that the new approach provides better solutions to these problems than previous predictivist accounts, and thus it constitutes a way forward in addressing both epistemic and pragmatic issues related to scientific prediction.

ACKNOWLEDGEMENTS

I am grateful to my supervisors Petri Ylikoski and Markus Lammenranta for their support during my doctoral studies and the writing of this dissertation. I wish to also thank Alkistis Elliot-Graves, whose research project on scientific prediction got me started on this project. The participants of the reading group "Prediction in a Changing Climate" at the University of Helsinki provided helpful feedback at the early stages of my research. Robert Northcott's and Richard Dawid's constructive comments and suggestions were of great help in finalizing the contents of the dissertation. I wish to thank all of them.

Helsinki, 26th October 2022
Pekka Syrjänen

CONTENTS

1 INTRODUCTION

"*Lights All Askew in the Heavens – Men of Science More or Less Agog Over Results of Eclipse Observations – Einstein Theory Triumphs*." So ran the headline in the New York Times on Monday November 10th 1919. The results of Arthur Eddington's experiment were in: light bends as predicted by Einstein's theory of general relativity. The paper inquired about the results from Sir Joseph Thomson, the President of the Royal Society in London. Thomson responded that the new discovery was the most important contribution to the laws of gravity since Newton, but stated that it was not possible to put Einstein's theory into intelligible words for the common person. However, he then added: "[w]hat is easily understandable … is that Einstein predicted the deflection of the starlight when it passes the sun, and the recent eclipse has provided a demonstration of the correctness of the prediction" (The New York Times 1919).

Thomson expressed a sentiment that is shared by many philosophers, scientists, and laypeople alike: *confirmed predictions* provide important evidence about the truth of scientific theories. In the philosophy of science, this view is broadly known as *predictivism*: predictions are special; they have a unique and powerful role in confirming scientific theories. The purpose of this study is to evaluate, and ultimately to develop, this view. *Do predictions, and what kind of predictions, justify belief in the truth of scientific theories?*

For the past 50 years, the standard answer to this question in the philosophy of science has been that scientific theories are worthy of belief when they make *novel* predictions, i.e. when they lead us to empirical results that were not used in their construction. This view has played a constitutive role in two perennial debates in the field: prediction vs. accommodation and scientific realism vs. anti-realism. The prediction vs. accommodation debate concerns the comparative epistemic value of novel prediction and accommodation (i.e. empirical results that were used in the construction of the theory): should predictions count for more, and if so for what reason? The scientific realism vs. anti-realism debate is about the fundamental nature of scientific knowledge: do scientific theories provide true descriptions of the world, or are they merely empirically adequate tools that are used for explanatory and predictive purposes? For many realists, the novel predictive success of scientific theories establishes realism as the superior option. These two debates, and the role of novel prediction in them, are going to occupy us in this study.

The claim that novel predictions have special epistemic value is backed by strong intuitions, and many prominent philosophers have endorsed it. The case for predictivism appears straightforward. If a theorist *accommodates* evidence in the construction of a theory, it is not surprising that the theory fits the evidence. If a theory is intentionally adjusted to fit some empirical results, this tells us more about the ingenuity of its creator rather than what the world is really like (see Worrall 1985, p. 323). However, if a theory leads us to some new results that were not foreseen or used in its

creation, it appears that the success needs now to be attributed to the theory itself. Especially, if the new predictions are surprising or impressive, this provides strong evidence that the theory must have somehow latched on to reality. Predictivist views along these lines have been advocated by philosophers such as Whewell (1840), Peirce (1883), Giere (1984), Maher (1988), Musgrave (1988), Psillos (1999), and Vickers (2013, 2019), and they have also been attributed to the likes of Descartes, Leibniz, and Bacon (see Musgrave 1974, pp. 1-2; Howson 1990, p. 225).

Despite its intuitive force, predictivism also faces strong challenges. These challenges have led many others to reject predictivism, embracing instead anti-predictivism. First, novelty is a contingent issue: whether some evidence was novelly predicted or accommodated is a historical matter that concerns the biography and intentions of the theorist. Yet, many philosophers of science have held that scientific confirmation must be a *logical* matter that depends on the contents of the theory and the evidence as such. After all, what more could be relevant to determining the degree of support for the theory than the contents of the theory, the evidence, background knowledge and assumptions, and the relationships between them, all of which are unaffected by the issue between novel prediction and accommodation? From the logical point of view, the question of novelty is wholly irrelevant (see Musgrave 1974, pp. 2-3). Second, predictivism appears incompatible with the highly plausible principle of evidential reasoning that we should use all the evidence that we have (see Hitchcock and Sober 2004, p. 6). From the predictivist point of view, *ignorance* becomes a virtue: it is better not to know of some evidence than to do know and use it. The problem with predictivism is not only that it is difficult to reconcile with these widely held epistemic principles, but that it seems directly to conflict with them. Those who have rejected predictivist views include Mill (1843), Keynes (1921), Hempel (1965), Horwich (1982), Howson (1990), Collins (1994), and Harker (2008).

The issue of predictivism has been called a paradox (see Barnes 2008). *On the one hand*, predictivism seems obviously correct. Surely, the best evidence for a scientific theory is its surprising, impressive novel success. For instance, how else could a theory such as general relativity predict phenomena such as the bending of light and gravitational waves unless it has latched on to reality? In contrast, once these phenomena are known, they could be incorporated into all kinds of theories, no matter how outlandish and out of touch with reality. *On the other hand*, how could our best knowledge of the world depend on contingent accidents? Is it really the case that if general relativity had been proposed at another time, so that phenomena such as the bending of light and gravitational waves had already been known in the scientific community and used by Einstein (or someone else) in the construction of this theory, we should now have less confidence in the theory, or not believe it at all? Surely, a serious theory of scientific confirmation should only cite factors relevant to the contents of the theory and the evidence and not be dependent on accidents of history.

This study evaluates predictivism in its various forms, attempting ultimately to resolve this paradox. We study the two central debates associated with predictivism, prediction vs. accommodation and scientific realism vs. anti-realism, and seek answers to the following questions: 1) do novel predictions provide an epistemic advantage over accommodations in science,

and 2) does novel prediction provide a compelling criterion of success for realist commitment to scientific theories? On both accounts, the answer that is reached here is (a qualified) 'no.' I argue that the focus on novelty has missed the mark on what matters to confirmation in science. Novel prediction provides, at best, modest advantages over accommodation under highly limited circumstances, but there is no general or important reason to prefer novel prediction to accommodation in science. Consequently, novel prediction is also not a criterion of success that scientific realists can (or should) use in the support of their position.

Based on the standard picture, rejecting the special value of novel prediction appears to place me squarely in the anti-predictivist camp. However, in the end, this may not be an appropriate description, because I also make the case that there is a (significant) grain of truth in predictivism. I argue that the predictivists have been correct in identifying 'prediction' as an important epistemic indicator in science; they have simply focused on the "wrong" kind of predictions, and understood prediction altogether too narrowly. Building on a previous approach developed by John Worrall (2006, 2009, 2014), I suggest that what matters in science is actually the *predictivity*, or predictive performance, of the scientific theory itself. What has gone wrong in predictivism is the *novelty*-part, and in particular the commitment to contingent accidents in confirmation theory. I introduce a new *predictive virtues approach*, which focuses on logical dimensions of the predictive performance of the theory itself. The new approach rejects the idea that scientific prediction should be understood narrowly in terms of the novel prediction vs. accommodation distinction and seeks to provide a new, more comprehensive framework for evaluating predictive success in science.

The study is organized as follows. In Chapter 2, I investigate the prediction versus accommodation problematic. The discussion is based on a novel classification scheme of the different type of approaches that philosophers of science have developed to justify the advantage of prediction over accommodation. I call these approaches 'methodological,' 'logical,' and 'agent-based,' based on the type of epistemic good that they identify as the source of the predictivist advantage. The methodological approach argues that the advantage of novel prediction stems from the methodological *dis*advantages of accommodation. It represents what I call a *negative* strategy in defending predictivism: the predictivist advantage is due to problems with its converse, accommodation. The agent-based approach argues that novel prediction provides an epistemic boost over accommodation because it speaks to the epistemic virtues of the theorist. This strategy is *positive*: it identifies a special benefit with prediction rather than a disadvantage with accommodation. Finally, the logical approach rejects the contingent novel prediction vs. accommodation distinction. It holds that scientific prediction should be understood from the point of view of the theory, and how the theory stands in relationship to evidence in the world. In the logical approach, a scientific theory makes logical predictions when it entails or implies empirical results (i.e. it has particular empirical consequences that follow from the postulates of the theory and appropriate auxiliaries) and it makes logical accommodations when empirical results need to be worked into the theory 'by hand' as they are (see Worrall 2006). In other words, what counts for confirmation is that the theory stands in the appropriate kind of logical or

structural relationship to empirical results, where the results follow from the theory rather than those results be built directly into the theory as they are. At the end of the chapter, I introduce anti-predictivist arguments, and explore what evidence from the sciences reveals about the prediction vs. accommodation issue.

Chapter 3 introduces the scientific realism vs. anti-realism debate, focusing on the role that novel prediction has played in it. The epistemic advantage of novel prediction over accommodation has often been defended with scientific realism in mind. In this chapter I discuss in more detail how the realist has pursued this strategy and what kind of challenges he or she faces from scientific anti-realists. At the heart of the realist's position is the positive 'No Miracles,' success-to-truth argument, according to which the best explanation for novel predictive success is the truth of the scientific theory. I introduce recent specifications to the No Miracles argument, including the emphasis on the surprisingness of the novel predictive successes and the selective realist strategy, where the realist has shifted from committing to scientific theories as a whole to committing instead to the theoretical constituents that were responsible for the novel successes of the theory. The anti-realist challenge to scientific realism is introduced and evaluated in relation to the novelty-based defense of scientific realism.

Chapter 4 begins the critique of the predictivist, 'novel predictions count for more' view. In this chapter, the negative approach to predictivism, where the advantage of prediction over accommodation is seen arising from methodological issues with accommodation, is evaluated in more detail. I discuss the distinct methodological disadvantages that have been attributed to accommodation, and argue that they fail to establish a reason to prefer novel prediction to accommodation in science. The problems that the methodological approach identifies are genuine, but when closer attention is paid to the respective problems of novel prediction, the epistemic asymmetry between prediction and accommodation either disappears or can be accounted for based on more fundamental epistemic principles than those related to prediction and accommodation. There are contexts where novel predictions provide some advantages, but even in these cases the advantage of prediction over accommodation is not very strong, and it can often be exceeded just by more accommodation.

Chapter 5 challenges the stronger, positive predictivist thesis according to which novel predictions provide the grounds for realist commitment to scientific theories. The previous chapter raised suspicions about the truth-conduciveness of novel predictive success, but what is yet missing is a direct challenge to the realist's success-to-truth inference, pursued on the realist's terms. This chapter develops such a challenge by building on van Fraassen's (1980, pp. 39-40) famous anti-realist argument, according to which the success of science can be explained by the harsh Darwinian competition to which scientific theories are subjected. I argue that contrary to realist responses (e.g. Musgrave 1988; Psillos 1999), this challenge, if developed more fully, works particularly against *novel* predictive success. The problem with the realist's success-to-truth inference is that it does not take into account the contingent context of scientific practice where novel successes are pursued. In a highly competitive context where many interdependent attempts at theorizing are made over a very long period of time, the

relationship between novel success and truth is a highly contingent and undependable one. In this, the advocates of logical theories of confirmation have been correct: contingency is the fundamental problem with predictivism. This chapter seeks to show more clearly why this is so.

Chapters 4 and 5 reject novelty-based predictivism in its various forms. This leaves the logical dissolution of the novel prediction versus accommodation issue on the table. In Chapter 6, I introduce the new (logical) predictive virtues approach, according to which scientific theories need to be evaluated based on their predictive performance. I explore four predictive virtues – veridicality, specificity, scope, and counterfactual depth – that probe in different ways the predictivity or predictive performance of a scientific theory, i.e. how well it performs in logically predicting empirical results. I argue that the predictive virtues approach is superior to previous attempts at going beyond the novel prediction versus accommodation distinction in that it is more discriminating between different degrees of scientific success and it can also handle problematic cases of scientific confirmation that appear challenging to previous predictivist approaches. It can also be of help to the scientific realist, who needs an alternative to novel prediction in identifying the type of success that makes theories worthy of realist commitment. I also discuss the relationship of the predictive virtues approach to other pertinent topics in the philosophy of science such as explanation and scientific progress.

The final chapter provides a conclusion, presenting some challenges and future topics of study. The predictive virtues approach opens up a number of new topics as well as new perspectives to old issues. Both the epistemic and pragmatic implications of each of the four predictive virtues introduced in the study require further study. The connections between the virtues also need to be further investigated. From the point of view of philosophy of science, what is most relevant is a new approach to predictivism and the defense of scientific realism. The use of predictive virtues in the defense of scientific realism and in developing the selective realist strategy warrants further investigation. Finally, there are also consequences for scientific practice. The results of the study speak, in particular, to the value of methodological transparency, replications (i.e. multiple tests), and the use of more data in scientific practice. There is also a more overarching consequence, as a "new" picture of scientific practice is supported, where building theories based on the evidence can be just as good as making bold conjectures (cf. Popper 1963).

2 PREDICTION VERSUS ACCOMMODATION

To evaluate the epistemic value of prediction in science, we first need to study the prediction versus accommodation debate. The question in this perennial debate is, do novel predictions provide an epistemic advantage over accommodation in science, and if so for what reason? This issue is now commonly framed in terms of what is called 'use-novelty' (see Worrall 2002): a theory novelly predicts empirical results when it fits results that were not used in its construction and it accommodates results that were used to construct it. However, there have been many other proposals about how to draw the relevant distinction between novel prediction and accommodation, as well as many different theories about why novel predictions should, or should not, count for more than accommodation. We shall go over these debates in what follows. Our goal in this chapter is to develop an overall understanding of predictivist and anti-predictivist arguments in the literature, and thus form the basis for evaluating predictivism in subsequent chapters.

A few clarifications concerning the nature of the debate are in order. First, the motivation. A central objective in the prediction versus accommodation debate has been to develop some more nuanced criteria of confirmation for scientific theories. Theories can obviously fit, or can be made to fit, all kinds of empirical results. For example, a (spurious) relationship between the theory and the evidence can be produced easily just by tacking irrelevant parts to the theory by conjunction (e.g. Schippers & Schurz 2020) or by making *ad hoc* adjustments (e.g. Worrall 2002, 2006), where some evidence is incorporated into the theory by modifying it in an unnatural way. In these cases, the empirical results arguably do not count much (if at all) in favor of the theory – let alone provide a compelling reason to believe that the theory is true. A central idea behind the prediction versus accommodation debate has been to identify the kind of empirical results that do provide genuine support for the theory. Novel predictions, it has seemed, might provide an example, even the best example, of the type of results that warrant strong belief in the theory. With novel predictive success, we can rule out many ways in which evidence may have been used in an inappropriate way (see section 2.3) and also gain strong, positive reasons to attribute truth to the theory (see Chapter 3). With a successful solution to the prediction versus accommodation problem in hand, philosophers of science could gain important insight into how genuine confirmation works in science.

Second, what does a successful solution of the prediction versus accommodation problem look like? There is disagreement over this question, but a good starting point is outlined by Mayo (2014, p. 80), who argues that the account should satisfy three desiderata: i) it should be objective rather than subjective or 'psychologistic,' ii) it should agree in important cases with the confirmatory evaluations of scientists, and iii) it should have a clear

epistemological rationale. The first desideratum means, roughly, that the account should not depend in some unexplained way on historico-biographical factors such as what evidence the theorist happened to be thinking about in private while developing their theory, but refer instead to some objective features of the scientific process (e.g. methodology, scientific tests, properties of the theory and the evidence). This desideratum is not accepted in all accounts of the prediction versus accommodation issue, but even in these accounts, it remains an important problem to explain why psychologistic facts do matter for confirmation (see, for example, White 2003, p. 670). The second desideratum has been considered central by just about all participants since the beginning of the contemporary debate on prediction versus accommodation. The debate has been pursued in a broadly naturalistic framework, where empirical evidence about the actual confirmatory judgments of scientists provides an important test for philosophical theories about scientific confirmation. Finally, a successful account of the issue should provide a compelling epistemological rationale for why novel prediction is better than accommodation. In other words, the account should explain and justify the (alleged) epistemic asymmetry between prediction and accommodation.

Third, a few notes on terminology. In what follows, 'prediction' and 'novel prediction' are short for 'use-novel prediction,' unless otherwise stated. 'Accommodation' or 'accommodated result' means the converse, i.e. an empirical result that fits a scientific theory but is not use-novel. There are other ways to define these words. In scientific practice, 'prediction' sometimes means prediction of the future, and sometimes it just means that a theory entails or implies some empirical consequences. The former kind of prediction is called 'forecasting' in this study, while the latter is identified as 'logical prediction.' To be a 'predictivist' is to hold that *novel* prediction provides special epistemic advantages over accommodation in science. (There is also a *logical* form of predictivism, but more on that later.) An 'anti-predictivist' believes that there are no special advantages to prediction. Also, I will generally speak of theories 'fitting' or 'implying' empirical results to make room for cases where the theory does not deductively entail the evidence (e.g. statistical hypotheses) (cf. Dellsén forthcoming). These words can be replaced by 'entails' in so far as we are considering a case where there is a deductive relationship between the theory and the evidence (with appropriate auxiliaries).

The chapter proceeds as follows. I begin by going over the early history of the prediction versus accommodation debate. The contemporary debate began to take shape in the 1970s and 80s, when important distinctions between different types of predictivist accounts started to emerge in the literature. I explain these distinctions, and discuss some stronger versions of predictivism that are now almost universally abandoned in the philosophy of science. Currently, the most popular approach to the advantage of novel prediction over accommodation is so-called *weak predictivism*. There are

many different types of weak predictivist theories, each of which is introduced in more detail. Where appropriate, some preliminary critique of these theories is also provided. Towards the end of the chapter, I discuss anti-predictivist challenges to predictivism, and consider briefly the question of what evidence from the sciences can tell us about the prediction versus accommodation issue. In the final section, I draw conclusions about where the discussion stands.

2.1 HISTORY AND TERMINOLOGY

The origin of the prediction versus accommodation debate is generally traced back into the 19th century, in particular to the Whewell–Mill debate on the inductive methodology of science. The Whewell–Mill exchange on prediction and accommodation itself was quite short. On the predictivist side, William Whewell (1840, II, p.230) argued that when theories predict facts that are of different kind than those that were contemplated in their formation, this compels us to believe that the truth of our theory is certain, because no accidents could produce such "an extraordinary coincidence." John Stuart Mill (1843, II, p. 23) took issue with this principle, arguing that novel predictions might well impress the ignorant masses, but scientific thinkers should place no value on the mere coincidence of the novelty of evidence. Mill's anti-predictivist view was probably first spelled out in more detail in the early 20th century by the economist John Maynard Keynes, who claimed that the virtue of prediction is "altogether imaginary" (see Keynes 1921, p. 349). Keynes pointed out that when theories are proposed prior to the testing of their predictions, there is typically already some grounds for proposing the theory. He argued that confirmation is based on the union of prior knowledge and the confirming instances. Whether the theory was developed first to make predictions or later to provide explanations makes no epistemic difference.

In contemporary philosophy, novel prediction re-emerged in a prominent role roughly a century after the Whewell–Mill debate, in the philosophy of science of Karl Popper. Popper (1963) famously argued that science and pseudoscience are demarcated based on the fact that scientific theories "stick their neck out" by making falsifiable predictions. Pseudoscientific theories, in contrast, can always be adjusted to explain (i.e. accommodate) any phenomenon in their domain. Pseudoscientific theories never run the risk of being falsified by the evidence, and so they are not supported by that evidence. On the face of it, Popper appeared to advocate a very strong predictivist position, where only predictions provide support for a scientific theory. However, it may not be apt to call Popper a predictivist in the sense that the term is used today, as for Popper even scientific theories were never confirmed by their successful predictions; the fact that they were able to resist falsification merely 'corroborated' them. Regardless, Popper introduced a distinction where predictions are superior to accommodations, and thus he

laid the groundwork for what was to follow in the philosophical debates on novel prediction.

The contemporary debate on the prediction versus accommodation problem started very much in consequence of the debates on Popper's philosophy of science. An issue that was recognized as important in improving Popper's demarcation criterion was to distinguish when modifications to scientific theories are genuinely predictive (i.e. falsifiable) versus ad hoc (i.e. unfalsifiable) (see Lakatos 1968, pp. 376-378). One of Popper's most prominent followers, Imre Lakatos, developed a solution: new scientific theories are falsifiable if and only if they have *excess empirical content* over their predecessors (see Lakatos 1970, pp. 116-118). This required that the new theory successfully predict a 'novel fact,' which according to Lakatos (ibid.) needed to be "improbable in the light of, or even forbidden" by previous knowledge. In this way, the first distinctive definition of novelty was introduced into the philosophical literature and applied as a criterion of success for scientific theories.

Lakatos's new criterion did not survive for long, as philosophers recognized that it was excessively strong.[1] The new criterion appeared to make the confirmation of scientific theories dependent purely on the *temporal* relationship between the theory and the evidence. As stated, Lakatos's criterion meant that any fact that was already recognized in the scientific literature could not confirm a new theory. This, as argued for instance by Elie Zahar (1973, p. 101), was clearly not in accord with scientific practice. A standard example is the general theory of relativity and its success in explaining the anomalous precession of Mercury's perihelion. The anomalous precession of Mercury's perihelion was known to the scientific community long before it was explained by general relativity, but it is considered to provide enormous confirmation for general relativity by all. The recognition of this defect in Lakatos's definition of a novel fact launched a new effort where philosophers of science sought to define the appropriate sense in which empirical results need to be novel versus accommodated so that they provide (strong) support for the theory.

Much of the groundwork for the contemporary debate on prediction versus accommodation was then laid out in the 1970s and 1980s. What started as a discussion about the appropriate definition of novelty thrived to such an extent that coming into the 1980s, the prediction versus accommodation issue emerged as a topic in its own right, as philosophers sought to develop theories about what it is about novel predictions that makes them superior to accommodations. During the 1980s, there was a lot of discussion about whether formal confirmation theories, namely Bayesianism, could incorporate predictivism. By the end of the decade, philosophers had defended all four possible positions on the issue: Bayesianism is 1) valid

[1] Even Lakatos himself walked back the criterion, observing that new theories could also explain old facts in a novel way. He argued that these would count as novel facts as long as the new research program could be rationally reconstructed as progressive (see Lakatos 1970, pp. 155-157).

because it supports predictivism, 2) valid because it does not support predictivism, 3) invalid because it supports predictivism, and 4) invalid because it does not support predictivism (see Brush 1994, p. 134). Around the same time, the contemporary scientific realism vs. anti-realism discussion also began to take shape, following the seminal work of van Fraassen (1980) and Laudan (1981). Seeking a response to van Fraassen's and Laudan's new challenges to scientific realism, realists were on the look for more stringent criteria of scientific success that could be used to defend realist commitment to scientific theories. Novel predictive success emerged as a promising candidate (see, in particular, Musgrave 1988), and thus novel prediction became a central concept also in the scientific realism vs. anti-realism debate.

Three developments took place in these first decades of the discussion that are particularly pertinent to understanding the contemporary state of the prediction versus accommodation debate. Due to these developments, contemporary predictivism has become predominantly 'local,' 'heuristic,' and 'weak.'[2] We shall next examine what is meant by these distinctions, and why local heuristic weak predictivism has emerged as the standard type of predictivism of the day.

The first development concerns *the scope* of the predictivist position. In Popper's writings, and later in Lakatos (1970), predictivism appeared in a global form: novel predictive success was presented as a universal standard for what counts as proper scientific success. However, in later discussion, it was recognized that the predictivist thesis could also be applied more narrowly, limited to more particular circumstances. We can thus draw a distinction between 'global' and 'local' form of predictivism, of which the former holds that prediction is always better than accommodation while the latter argues that the predictivist advantage comes into effect only in certain situations (see Hitchcock and Sober 2004).

The local version of predictivism overcame global predictivism through both theoretical development (e.g. Maher 1988) and more detailed attention to case examples (e.g. Howson 1988). In particular, it was recognized that sometimes accommodation of evidence is perfectly acceptable, and in no way inferior to prediction from the confirmatory point of view. Howson (1988, pp. 387-388) provides the following simple example. Consider that we have a box that contains red and black tickets in a proportion p of red (r) to black (b). To evaluate the parameter p, we draw tickets from the box, concluding that p = r/b. Here, the evidence is used in evaluating p, but it is clear that it (strongly) confirms our estimate. Similar examples abound. For instance, in scientific polling, estimates are drawn directly based on the data, but this can nonetheless be a reliable way to produce good estimates (see Mayo 1996, pp. 272-273). Another example might be a detective who uses evidence from a crime scene to solve a murder mystery. DNA found on the crime scene appears

[2] The terminological distinctions between different types of predictivism that form the basis of the discussion in what follows were introduced by Hitchcock and Sober (2004, pp. 3-4). All forms of predictivism that they identify had emerged in the literature by the end of the 1980s.

perfectly admissible as a source of evidence that can simultaneously generate and provide support for a hypothesis about who committed the murder.

Prior to the recognition of such examples, some attempts were made at developing a general argument for global predictivism. Such an argument could be attempted in the following way: novel prediction is always better, because only in cases of novel prediction empirical facts have a chance of falsifying the hypothesis (see Giere 1984, p. 161). In cases of accommodation, there is no chance for falsification, as the theory was developed on the condition that it fit the evidence. Howson (1990, p. 229) provides a compelling response, pointing out that this argument contains a fatal flaw: facts, being what they are, never have a 'chance' of falsifying anything. They either falsify or they do not, regardless of whether a hypothesis was designed to fit them.[3] Thus, the argument for global predictivism collapses. Just about all contemporary predictivists have agreed that local predictivism is a more promising way forward, and sought to develop accounts where the predictivist advantage is circumscribed in some appropriate way.[4]

The second development concerns the appropriate definition of novelty. Musgrave (1974) laid out the menu of options that is used largely even today. Musgrave distinguishes between three competing definitions of novelty: 'temporal,' 'heuristic,' and 'theoretical.' The temporal definition says that evidence is novel if it was not known to science at all when the theory was proposed (cf. Lakatos 1970). The heuristic definition, which later became known as the 'use-novelty' definition (see Worrall 2002), says that evidence is novel if it was not used in the construction of the theory. Finally, the theoretical definition states that evidence is novel if it cannot be accounted for by any of the theory's existing rivals.

According to Douglas and Magnus (2013, p. 581), the heuristic, use-novelty view became popular because philosophers of science wanted to maintain an important concept of novel prediction while making room for emerging evidence about scientific theory evaluation and philosophical intuitions about scientific confirmation. The temporal view of novelty, as noted before, appeared excessively restrictive. There are clear examples in the scientific literature where known evidence is considered strongly to confirm a new theory (e.g. general theory of relativity). In terms of philosophical intuitions, the temporal view also lacked a compelling epistemic rationale, as it is hard to see why the mere fact of the temporal order in which the theory and the evidence are introduced should matter for confirmation (e.g. Worrall 2002, p. 194). Consider, for example, that a theorist proposes a new theory completely unaware of some evidence e, but it just so happens that e was

[3] The crux of this response may be easiest to see in the following way. Consider that an experiment E is performed where an empirical result e is discovered, and a hypothesis H is then designed to fit e. Later, it is discovered that E was flawed, and actually ~e holds. H is refuted despite being designed to fit e.

[4] An important exception is Worrall's (2006) logical predictivism, which does hold that there is a certain kind of global advantage to prediction. However, Worrall defines prediction and accommodation in a very different way to the standard, contingent use-novelty definition (see section 2.4).

discovered earlier by another scientist and is just about to be published in the literature. In other words, e is not temporally novel, but the theorist had no way of knowing about this evidence, let alone using it in constructing his theory. This mere historical accident, it seems, could not possibly affect the degree to which e supports the theory. Thus, it was suggested that the underlying factor that matters for confirmation, if indeed something does, concerns the *use* of evidence: evidence is novel if it was not used by the theorist in the construction of the theory.

The other alternative, theoretical novelty, was advocated by Musgrave (1974) himself. However, this option has not enjoyed much support among contemporary predictivists, as it does not appear to get at the heart of the issue: theoretical novelty either lacks a compelling epistemological rationale, or it turns predictivism into a trivial truth. Worrall (2006, pp. 36-37) notes that it is difficult to find a rationale for why new theories could not be supported as strongly by empirical facts already entailed by older theories as these facts support the older theories. If a new theory entails an empirical fact that happens also to be entailed by another theory, but this fact played no part in the formulation of the new theory, there seems to be no reason why the fact should fail to also support the new theory. Ladyman (1999) makes a similar argument, objecting that the idea of theoretical novelty makes scientific confirmation dependent on the contingent accident of which theory comes first: the first theory to entail the theoretically novel fact gets full credit, while the ones that follow no longer cannot.[5] Finally, Barnes (2008, p. 2) points out that if theoretical novelty is understood in the unobjectionable sense that a theory which explains an empirical fact that no other theory at the time is able to gets (temporary) credit over its rivals, this makes predictivism into a trivial truth, as theories that have more evidence are obviously more supported than theories that have less evidence. Such observations have made theoretical novelty a less attractive option in capturing the epistemically relevant notion of novel prediction in science.

The use-novelty definition, even if relatively more promising than the alternatives, does come with problems too. From the point of view of use-novelty, confirmation appears to become a person-relative matter, as whether some evidence was used in the construction of the theory or not concerns the actions and intentions of the theorist (see Musgrave 1974, pp. 12-15). This is also a puzzling consequence, at least from the point of view of logical theories of confirmation, which argue that theory confirmation depends only on the

[5] Ladyman (1999) responds to Leplin (1997), who argues for an account of novelty which combines a version of use-novelty and a strong version of theoretical novelty, theoretical uniqueness, which requires for the theory to be the *only* predictor of the use-novel fact. Leplin's motivation for introducing the theoretical uniqueness condition arises from the desire to defend scientific realism. He argues that if there is more than one theoretical way of achieving a use-novel prediction of some fact, it is demonstrated that it is possible to achieve this success by chance, which means that a truth-based explanation is no longer required. The argument is formulated with the familiar problem of underdetermination in mind, which Leplin seeks to resolve by appealing to theoretical uniqueness (see Chapter 3 for further discussion on scientific realism and its challenges).

contents of the theory and the evidence as such. Furthermore, whereas temporal novelty and theoretical novelty are relatively simple to identify in practice, use-novelty may not be, because it seems to require information about the private thoughts of individual scientists. Such problems prompted for some time debate about alternative versions of the use-novelty criterion.[6] However, a simple use-novelty definition ultimately prevailed, following other developments that rejected the stronger, global versions of predictivism. Douglas and Magnus (2013, p. 581) explain that given that accommodation was also recognized to have epistemic value, it was no longer paramount to provide an exact definition that could be clearly evaluated in all cases. As far as there were questions about whether some evidence was used in a particular case or not, it could simply be assumed that an accommodation was performed, which would also count for something. Useful philosophical analyses could nonetheless be conducted about the myriad of cases where it is (relatively) clear that evidence was not used by the theorist. Later predictivist theories have also continued to provide various solutions to the problem of why the use of evidence is epistemically relevant (see sections 2.3 and 2.5).

Despite use-novelty emerging as the leading view of how to define novelty, it is useful to recognize that use-novelty is not mutually exclusive with either temporal novelty or theoretical novelty. An empirical fact could be use-novel and also theoretically and/or temporally novel. Particularly, in the case of temporal novelty, it can be pointed out that even if temporal novelty does not count as such, temporal novelty could be a uniquely clear way to recognize use-novelty: empirical facts that were not known to anyone could not have been used by the theorist in the construction of the theory, thus guaranteeing use-novelty. Thus, even if one recognizes the use-novelty criterion as the most fundamental one from the epistemic point of view, more nuanced positions remain on the table, where other types of novelty could also count for some extra confirmation in virtue of what they reveal about use-novelty.

The final important development in the prediction versus accommodation debate concerns the nature of the predictivist position. Hitchcock and Sober (2004) introduced a distinction between 'strong' and 'weak' predictivism. Strong predictivism claims that there is *an inherent* advantage to prediction over accommodation. This means, roughly, that there is something about prediction as such that makes it epistemically advantageous to accommodation. In contrast, weak predictivism sees the predictivist advantage as *an indirect* one. It maintains that prediction is better because it tracks or is symptomatic of some other feature that is of epistemic import.

Strong predictivism has been defended most prominently through a simple inference to the best explanation argument ('IBE'). This argument appeals to the different explanations that seem to be called for in cases of novel

[6] See Gardner (1982) for discussion. He proposes that evidence is novel if it was not *known* to the person who constructed the theory.

versus accommodative success. Novel predictive success, it seems, can only be explained by the truth of the theory, as it is highly unlikely that a false theory should happen to entail something real that was unknown (or not used) before (cf. Whewell 1840). In contrast, in case of accommodative success, there is another explanation available: the theorist, knowing the evidence in advance, made sure to design his theory so that it entails the evidence. In other words, the truth of the theory needs to be invoked to explain the success of the theory in case of novel success, but in the case of accommodative success, we can simply point to the theorist's intention of accommodating the evidence. Thus, there is an inherent, explanatory advantage to prediction over accommodation.

This argument is now almost universally abandoned in the prediction versus accommodation debate. (It does still appear in the literature on scientific realism, a point which warrants some discussion later on.) A number of philosophers on different sides of the issue have argued that this argument does not establish an epistemic difference between novel prediction and accommodation (e.g. Horwich 1982, pp. 111-117, Collins 1994; Barnes 2002a; White 2003). This is because no reason has been given for why the design-based explanation that is appealed to in the case of accommodation *competes* with the truth-based explanation that is invoked in the case of prediction. These explanations refer to fundamentally different subject matters, the one being about the theorist (accommodation and design) and the other about the theory (prediction and truth). In order to find a valid reason to prefer novel prediction to accommodation, some *explanation* (or justification) needs to be given for why the distinction of whether a theorist used or did not use some evidence is relevant to the degree to which we should believe that the theory is true. Weak predictivism naturally follows: philosophers have attempted to find the epistemic good, feature, or property that use-novel prediction speaks to that establishes the predictivist advantage.

To summarize, contemporary predictivists now commonly advocate predictivism in the local, heuristic, and weak form. The predictivist recognizes that accommodation has confirmatory value, but novel prediction is (typically) more powerful. What counts in deciding whether an empirical result was predicted or accommodated is whether the theorist used that result in the construction of the theory. Facts about which evidence was or was not used by the theorist are not always unambiguous, but in many cases the distinction is clear enough to make judgments about scientific confirmation. Finally, prediction is not understood to be better in any inherent sense, but because it correlates with some other feature or property that carries epistemic import.

2.2 INTRODUCTION TO WEAK PREDICTIVISM

The main differences in contemporary predictivist accounts concern the details of the weak predictivist argument. There is now a motley of weak predictivist theories that put forward very different ideas about the epistemic good or good-making property that novel prediction is associated with (see Barnes 2018). These are examined in what follows (sections 2.3, 2.4, and 2.5).

It is useful to start by categorizing different approaches within the weak predictivist position. I propose the following categorization scheme, which is based on the type of epistemic good or good-making feature these theories associate with prediction. *Methodological theories* focus on scientific methodology and methodological choices involved in generating or testing scientific theories. These theories argue that the predictivist advantage emerges due to particular methodological problems with accommodation rather than any specific advantage to novel prediction per se. The basic idea is that accommodations are associated with certain less reliable methodologies, and therefore prediction provides stronger confirmation. *Agent-based theories* hold that the important epistemic good behind the prediction versus accommodation issue is the 'reliability' of the theorist who constructed the theory, i.e. how probable it is that the theorist has generated a true rather than false theory. Prediction is more valuable than accommodation because it provides evidence that the theorist has been a reliable source. Finally, *logical theories* argue that the important underlying feature in the predictivism debate concerns the logical or structural relationship between the theory and the evidence. The logical approach abandons the standard use-novelty criterion, rejecting the idea that contingent factors such as who came up with the theory, when, and for what reason should have a place in confirmation theory. Instead, it argues that what matters is the *end product*, i.e. the theory, and its relationship with the evidence, as such. The basic idea is that a certain type of use-novel evidence, defined now roughly as evidence that is not needed in the construction of the theory, supports the overall theory, whereas accommodated evidence that is specifically needed in the construction of the theory has less epistemic force.

These approaches come at the issue from very different perspectives. Two approaches hold that there is an epistemic advantage to standard (contingent) use-novel prediction: the methodological approach and the agent-based approach. However, they argue for the predictivist advantage in a very different way. Agent-based theories adopt an *explanatory* strategy to derive a predictivist advantage. These theories hold that only predictive success calls for a reliability-based explanation. This provides a good reason to increase confidence in the theory, because theories that come from a reliable source are more likely to be true. The agent-based approach has been developed chiefly as a solution to the problem of the explanatory argument for strong predictivism (see White 2003; Barnes 2008). 'Truth' and 'the theorist's use of evidence,' as was noted in the previous section, do not appear to provide

competing explanation for why a theory successfully entails some evidence. Referring to the theorist's reliability in selecting the theory potentially solves this problem: the best explanation for the emergence of an empirically successful theory could be that it was selected by the theorist on the condition that it entails some evidence (accommodation), or that it was selected by the theorist in some reliable way (prediction).

In contrast to the agent-based approach, the methodological approach makes use of a certain type of *causal* argument. These theories argue that accommodation causes, or is associated with something that causes, for the theory construction, selection, or testing process to become more *un*reliable (no matter how reliable or unreliable the theorist might otherwise be). In other words, the methodological approach turns the discussion on its head: the issue concerns the side where theorists do use evidence (accommodation). For an advocate of the methodological approach, the predictivist advantage is basically in that novel prediction enables evaluators of the theory to allay concerns about the problematic consequences of accommodation.

Finally, the logical approach denies that there is room for historical, contingent considerations in confirmation theory. In this, it dissolves the original novel prediction versus accommodation distinction, so that novelty no longer plays an epistemic role. We thus come back to logical theories of confirmation, where epistemic support for the theory depends on the relationship between the theory and the evidence as such (along with relevant background knowledge and assumptions). However, equipped with predictivist ideas, we come back with a twist: in the logical predictivist approach, predictivist insights are incorporated into logical confirmation theory to capture more nuance in the confirmatory relationship between the theory and the evidence. The logical approach to predictivism has been developed most of all by John Worrall (see, in particular, Worrall 2006), whose theory is discussed in more detail in sections 2.4 and 6.1.

Despite their differences, or perhaps in virtue of their differences, these approaches are not always incompatible. In particular, the methodological and the agent-based approaches could be seen as providing complementary solutions to the predictivism issue. This is basically argued by Barnes (2008), in his defense of an agent-based solution (although Barnes does not use the same terminology as I do, calling these 'virtuous' and 'unvirtuous' predictivism instead). The logical approach, at least if taken strictly by the letter, is incompatible with both the methodological and the agent-based approaches. However, it too has been developed into a version that incorporates some of the insights of the methodological approach (see Schurz 2014). The variety of options on the table gives the weak predictivists a significant degree of leeway in defending their position.

A final general point to highlight about the weak predictivist approaches pertains to their application. The weak predictivist theories address the issue of theory confirmation from the point of view of *an evaluator*

of a scientific theory (i.e. somebody who evaluates the degree to which they should believe the theory). In order to decide whether and to what degree he or she should believe a particular theory, the evaluator looks at the theory and the evidence, and comes up with an estimate based on the contents of the theory, the evidence, and background knowledge. The question then becomes, upon learning that some evidence was predicted rather accommodated, should the evaluator revise their beliefs about the theory? The weak predictivist argues that the answer is a qualified 'yes:' if the evaluator is not in a position to directly evaluate the epistemic good or good-making feature that prediction speaks to, the predictivist advantage follows. Based on the predictive success, the evaluator is able to conclude that the epistemic good or good-making feature is likely there, and thereby increase their confidence in the theory.[7] These epistemic goods, and the weak predictivist approaches that are based on them, are examined next.

2.3 THE METHODOLOGICAL APPROACH

There are three theories in the literature that adopt the methodological approach to weak predictivism: Hitchcock and Sober (2004), Lipton (2004), and Mayo (1996). Each of these theories identifies a particular problem with accommodation, and argues that in cases where this problem cannot be addressed directly, a predictivist advantage follows for scientific evaluators.

Hitchcock and Sober (2004) develop a methodological argument that applies to curve fitting in statistics. The setup is as follows. Scientists have collected a dataset (D) from a target population and seek to model a relationship between two variables (X and Y) in the data by fitting a curve into the dataset – a standard problem in statistics. The scientists are looking for a curve that optimizes predictive performance with further data points outside the sample, i.e. a curve that is maximally predictively accurate with regard to the target population. The problem is that the sample data, as data always, are 'noisy:' the data capture idiosyncratic features of the sampled population as well some degree of measurement error. To find the optimal curve, the scientists cannot therefore, for example, just draw a curve that goes through each data point in the sample data. Such a curve would be maximally accurate *within* the dataset, but its predictive accuracy outside the sample would be poor, as it has now accommodated all of the noise in the dataset. In other

[7] Barnes (2008) further distinguishes between 'tempered' and 'thin' predictivism. Tempered predictivism holds that information about prediction and accommodation carries epistemic import for scientific evaluators (i.e. the relevant epistemic goods or good-making features underlying prediction and accommodation are not always evaluable directly in practice, and thus prediction matters). This is the type of predictivism that has been mostly at issue in the prediction versus accommodation debate: does the information that some evidence was predicted rather than accommodated carry epistemic force for scientific evaluators? However, it is also possible to consider a more modest, thin version of predictivism, where prediction *correlates* with an epistemic good that *is* evaluated directly. Given that the epistemic good is evaluated as such, prediction itself is not used for epistemic purposes in these cases.

words, it *overfits* the data. To achieve optimal predictive accuracy outside the sample dataset, a simpler curve is needed that is able to avoid overfitting the data.

How can the optimal curve be selected so that overfitting is avoided? Hitchcock and Sober appeal to Akaike's (1973) famous result, according to which an unbiased estimate of a model's predictive accuracy can be obtained by assessing both its fit-to-data and complexity (where complexity is measured by the number of adjustable parameters that the model contains). Akaike showed that the estimate for model M ≈ log[Pr(Data|L(M)] – k, where L(M) is the likeliest member of M, and k is the number of adjustable parameters in M. This is known as the Akaike Information Criterion (AIC), which is used to 'score' models based on their expected predictive accuracy. Basically, the criterion measures the model's fit-to-data, but punishes it for complexity (hence, 'minus k'). Given that overfitting is a problem of excess complexity, the AIC can be used to combat overfitting in curve fitting.[8]

With this set as background, Hitchcock and Sober produce a nuanced predictivist argument by considering five possible cases where we evaluate competing models of the relationship between the two variables (X and Y) in the dataset (D). In all cases, we consider a prediction competition between two modelers, "Penny" the predictor and "Annie" the accommodator. The goal of the competition is predictive accuracy with out-of-sample data. We are given the following information about Penny and Annie. Penny constructed her model, M_p, as follows: she first obtained an initial fragment of the dataset, D_1, and fit a curve to this fragment of the dataset. She then used M_p, to predict the remaining data, D_2, to a high degree of accuracy. Annie, on the other hand, simply used all of the data points in D to fit her model, M_a. In other words, she accommodated all of the data. The question then becomes, should we prefer M_p to M_a with regard to expected predictive accuracy?

The cases considered by Hitchcock and Sober, and their respective responses to these cases, are as follows. 1) We do not know either the model Penny or Annie has proposed, but we do know that it happens to be the same exact model in both cases. In this case, there is no reason to prefer Penny's model: given that the model is the same, its predictive performance is going to be equal in both cases. 2) We know both Penny's and Annie's models, and are able to calculate their AIC scores. We should prefer the model with the higher AIC score, given that it has the higher expected predictive accuracy. The information about the method by which the models were constructed becomes irrelevant. 3) We do not know either Penny's or Annie's model, but we know that Annie fit D exactly while Penny fit D_1 exactly and then proceeded to accurately predict D_2. Here, both are likely to be guilty of overfitting, but Penny has provided herself with some extra protections, as she could not have overfit data that she did not yet have (i.e. D_2). Penny's model should be preferred, as

[8] There are other curve selection methods that aim for the same goal, e.g. cross-validation. We will return to these methods more closely in Chapter 4.

it is likely to have the higher AIC score. 4) Both Penny and Annie choose the best-fitting curve based on the highest AIC-score, but Penny uses only D_1 while Annie uses all of D. *Annie's* model should be preferred, as Penny could not have chosen a curve with a better AIC score by using only a part of the data. In other words, accommodation is better once safeguards are in place to combat overfitting. 5) Penny and Annie construct their models as before (Penny uses D_1 while Annie uses all of D), but we do not know what method they used for curve selection. Penny's predictive performance with data D_2 provides a reason to believe that she has avoided overfitting. In Annie's case, we do not have information about overfitting one way or the other. Penny's model should be preferred, as there is evidence that Penny has avoided overfitting the data.

Hitchcock and Sober identify two cases in which prediction is better than accommodation: in cases where we cannot evaluate models or hypotheses directly and either (i) have reason to suspect overfitting may have occurred (case 3) or (ii) do not know whether or not appropriate measures have been taken to avoid overfitting (case 5). Interestingly, Hitchcock and Sober argue that out of all the options on the table, the best one is the model that accommodates all of the data while AIC is used to combat overfitting (case 4). In other words, they argue that it is best to use all of the evidence as long as the danger of overfitting is addressed. Notably, Hitchcock and Sober do not make any assumptions about how common each of these type of cases is likely to be in scientific practice. Their only concern is to demonstrate how a predictivist advantage can sometimes arise; for all that has been said, all cases could be like case 4 in science (ibid. p. 21, fn. 23).

The account Hitchcock and Sober provide is a distinct version of methodological weak (heuristic, local) predictivism. Their account holds that prediction can sometimes be better because accommodation is associated with a certain methodological flaw (overfitting). The underlying argument is basically an inductive one: provided that we are interested in predictive accuracy, and there is a known obstacle to achieving predictive accuracy (overfitting), it is better to go with a model or hypothesis that has demonstrated some ability to overcome this obstacle than one that has not (see Hitchcock and Sober 2004, pp. 19-20). We can then hope that the world continues to cooperate and predictive success continues with more out-of-sample data.[9]

Another methodological weak predictivist theory is provided by Lipton (2004, pp. 164-183). Similarly to Hitchcock and Sober (2004), Lipton identifies a distinct methodological problem with accommodation, and then proceeds to argue that novel prediction should be preferred in science.

[9] Hitchcock and Sober have been criticized both from the perspective that their predictivist advantage holds only in even more circumscribed cases than they have argued (see Lee 2013) as well as that it actually holds in a (much) wider range of cases (see Douglas & Magnus 2013, pp. 582-584). In particular, Douglas and Magnus argue that there may often be uncertainty about whether the assumptions underlying AIC hold in practice. For this reason, they hold that *temporally* novel prediction can provide even further assurances that overfitting has not occurred. I will consider this argument in Chapter 4, where different aspects of the methodological approach to weak predictivism are evaluated.

However, the problem he identifies applies more widely in different theory generation contexts. Lipton argues that the problem with accommodation is that if theorists develop their theory with the purpose of accommodating for some empirical results, they may, intentionally or inadvertently, 'fudge' their theory or their auxiliaries in an effort to find a theory that entails the empirical result. Fudging means, roughly, that the theorist makes some unnatural adjustments or modifications to their theory or auxiliary assumptions. Lipton explains that in case of the theory, this has costs in terms of theoretical virtues (namely, simplicity) and in case of the auxiliaries, it has costs in terms of epistemic relevance. In both cases, the result is an inferior theoretical explanation of the evidence that warrants less confidence.

The crucial part of the argument that makes Lipton a methodological weak predictivist concerns the application of his account. Even though Lipton appeals to properties of the theory as the underlying issue with accommodation (i.e. lack of simplicity), he explicitly denies that we can reduce the prediction versus accommodation problem to the logical or structural properties of the theory and the evidence (cf. section 2.4). He argues that even with the theory and the evidence fully public and available for scientific evaluation, the disadvantage of accommodation is there (if not always at least in many cases). He explains that this is because the kind of inductive support that is available in science is "translucent, not transparent" (ibid. p. 178). Simply put, scientists are not perfect judges of the degree of fudging that has gone into constructing a theory. Even if a scientist is directly in possession of the relevant evidence and can seek to evaluate specific virtues of the theory, it may still be that problematic fudging has taken place in cases of accommodation. Thus, the *act* of accommodation becomes the epistemic issue for the scientist: learning that evidence was accommodated rather than predicted should lead evaluators to lower their confidence in the theory.

Lipton makes a few additional specifications. He clarifies that not all cases are equal in terms of concerns about fudging. The problem is more or less severe depending on the complexity of the situation. Concerns about fudging are strongest with complex high-level theories and they are either reduced or disappear altogether in the case of simple empirical generalizations. As an example of the former, Lipton (ibid. p. 172) cites the theory of relativity and its use-novel prediction of the anomalous perihelion of Mercury. Lipton argues that predicting this empirical result was better than accommodating it, because in such a complex situation, accommodation would have raised concerns about fudging. As an example of the latter, we can take any simple empirical generalization (e.g. 'all sparrows employ a particular courtship song'). In these cases, Lipton holds that it hardly matters at which point in collecting the evidence the generalization is proposed; its degree of support depends on the evidence as such. In this way, we can see a continuum from simple to complex or unclear situations, where the confirmatory asymmetry between prediction and accommodation increases the more complex the situation becomes.

An important limitation of Lipton's account is that he provides no precise criteria or characterizations for how degrees of fudging can be compared between different theoretical systems (or, indeed, what fudging actually "looks" like). Lipton concedes that he is unable to do so, arguing that his position does not hinge on providing such precise criteria. He writes (see Lipton 2004, p. 180): "It is enough that we have good reason to believe that different theoretical systems enjoy different degrees of support from data they all fit, and that the requirements of accommodation may be in tension with the desire to produce a highly confirmable theoretical system." In light of this, we may view Lipton's argument at least partly as an intuitive one, where the idea of fudging is expected to be clear enough based on the overall description provided. Another question that could be raised about Lipton's account, which is also related to the first point, concerns the claim that scientist are unable to evaluate fudging as such. Lipton does not show how fudging dupes scientists in practice, or provide evidence that it does. The question of whether or not we can expect scientists to be generally misled by fudging will require further investigation on our part (see Chapter 4).

The third and final methodological theory on the prediction versus accommodation issue has been developed in multiple publications by Mayo (e.g. Mayo 1996, 2009, 2014). Mayo's theory is actually developed as a solution to what she views as shortcomings of the traditional prediction versus accommodation distinction, so her account as a whole thus represents a dissolution of the prediction versus accommodation issue rather than a specific predictivist theory. However, in certain scientific contexts, Mayo identifies methodological problems that specifically concern accommodation, and argues that in these cases novel prediction is the solution. In this, she makes potential contributions to the methodological case for weak predictivism that we need to study more closely.

First, a few words about Mayo's overall confirmation theory. Mayo argues that the underlying issue behind the preference for use-novel prediction is actually a deeper concern about the *severity of the test* that the empirical evidence (and the methodology used in the test) provides to the theory. She points out that the alleged problem with using evidence in the construction of a theory is that this is "easy:" if the theorist is willing to make enough adjustments, a successful fit is ultimately assured no matter what the data are (cf. Lipton 2004, pp. 164-183). However, Mayo notes that it is actually a different thing for a hypothesis to pass a test no matter what the data are versus for a hypothesis to pass the test *given that the hypothesis is false*. According to Mayo, what matters for confirmation is the latter rather than the former, i.e. that the theory passes a severe test that probes the probability that the theory is in error.

Mayo argues that the test severity criterion and the use-novelty criterion do not always coincide; but, in these cases, what counts for confirmation is test severity rather than use-novelty. Mayo argues that it is possible for a theory to depend fully on evidence that is used in its

construction, but the evidence still provides a maximally severe test to the theory. A simple example is inferring the average SAT score of students in the class. Using the students' SAT scores to calculate the average guarantees that the hypothesis passes no matter what the data are, but test severity is nonetheless maximized. Mayo holds that there are several cases in science where this principle applies, so that data that are used in the construction of the theory nonetheless test the theory to different degrees of severity (e.g. inferring statistically significant differences, estimating or measuring parameters, accounting for anomalies) (see Mayo 2014, p. 82). In each such case, Mayo holds that the degree of confirmation depends on test severity rather than whether or not the evidence was used in the construction of the theory.

We will return to consider Mayo's overall confirmation theory briefly again in Chapter 6. In the current context, what is most pertinent is Mayo's weak predictivist argument, which is embedded within her overall approach. Mayo discusses one important situation where accommodation is indeed the problem (from the point of view of test severity) and where novel prediction in particular is the solution (see Mayo 1996, pp. 294-318). This situation arises in the context of statistical hypothesis testing, in cases where hypothesis construction or modification occurs after the results of the statistical test are known. In classical hypothesis testing, a hypothesis is proposed in advance, and it is then tested against the data with a predesignated threshold of probability for error (often, $p = 0.05$). When the hypothesis is constructed before the statistical test is performed, it is tested at the appropriate, predesignated level of statistical significance. However, if the hypothesis is constructed *later* based on the data itself, Mayo argues that a problem arises: if the scientists go snooping through the data looking for multiple relationships between variables, the higher the probability is going to be that they will uncover *some* result that has passed the statistical threshold by chance. She argues that in such a case the actual severity of the test is much lower than what is implied by the select threshold. By searching through multiple possibilities, the scientists allowed for many opportunities for a hypothesis to pass merely by chance. Thus, there may be a significant probability that a hypothesis that has been selected via such a method is indeed merely successful due to chance, and such hypotheses warrant lower confidence than predicted hypotheses due to the way the data has been used in their construction.

Due to the hypothesis hunting problem, Mayo endorses a predictivist stance in the context of statistical hypothesis tests: after-the-fact construction of hypotheses alters the statistical significance level of the test, so pre-trial hypothesis designation (i.e. novel prediction) should be preferred. Mayo clarifies that the predictivist stance can be understood in a more nuanced way based on her overall approach to confirmation. What ultimately matters is that the level of severity of the test is reported accurately, whether prediction or accommodation was used. She notes that although prediction is

often better in practice, accommodation need not always be worse. For example, a large dataset with a low number of variables could allow an accommodated hypothesis to be constructed in a reliable way, and Mayo leaves open that scientists can be creative in looking for other ways to make tests more reliable (e.g. by using other methods or arguments that violate the use-novelty criterion while nonetheless meeting the test severity criterion).

To summarize, we have now introduced three methodological weak predictivist arguments that highlight the *dis*advantages of accommodation. Each argument provides a unique answer to the puzzle of why the use of evidence is relevant to the degree to which we should have confidence in the theory: when theorists do use evidence, they may end up making various 'bad' methodological choices that lower the probability that the theory at hand is true (or empirically successful). Methodological weak predictivists argue that the use of evidence can lead to overfitting, fudging, or hypothesis hunting. For this reason, we should prefer novel prediction to accommodation in science. Novel prediction provides evidence that neither overfitting nor fudging nor hypothesis hunting has taken place, and so we are able to retain unrevised support for the theory (i.e. epistemic support depends on the theory and the evidence as they have been presented to us). Whether these arguments are able to justify a preference for novel prediction over accommodation in science is evaluated in Chapter 4.

2.4 THE LOGICAL APPROACH

The second approach to weak predictivism on our list is the logical approach. In the logical approach, novel prediction no longer plays a confirmatory role as such, so we have in hand an approach that challenges the original prediction versus accommodation distinction. However, as formulated in the context of the prediction versus accommodation debate, this approach is still very much rooted in the original use-novelty criterion. The approach has been developed most of all by Worrall, whose mature logical account was made clear in Worrall (2006). We will focus predominantly on Worrall's work, and the philosophical commentary to it, in what follows.

Worrall has been a central participant in the prediction versus accommodation debate since the 1970s. In early publications (e.g. Worrall 1985, 1989b), Worrall defended the standard use-novelty criterion of success (developed originally by himself and Zahar), according to which scientific theories gain epistemic support from the empirical results that were not used in their construction. The problem in using empirical results in the construction of the theory, argued Worrall (1985, pp. 323-324), is that an after-the-fact adjustment to the theory tells us more about "the ingenuity of man" rather than the actual blueprint of the Universe. With the results already known, there may be many ways to bring them into agreement with the theory, and it is up to the ingenuity of the theorist to make use of those opportunities.

Use-novel prediction, in contrast, generates some intuitive warrant for thinking that the success of the theory must have some other, more substantive explanation (e.g. truth). For this reason, use-novel prediction should be preferred to accommodation in science.

Already at this time, Worrall was focused on an important problem faced by the use-novelty account: it appears difficult to evaluate in practice whether some evidence was or was not used in the construction of the theory (see Worrall 1985, pp. 310-311). Worrall sought to mitigate this problem by identifying different types of cases where the proper evaluation could be conducted relatively easily (see ibid. pp. 311-318). He argued that the fact that evidence has been used in the construction of the theory can be recognized at least in cases of 'exception incorporation' and 'parameter adjustment,' and in contexts where providing an explanation for the empirical successes of an older theory is used as a litmus test for a new theory. The first case refers to situations where a generalization is adjusted based on an empirically discovered anomaly or exception, the second to cases where a theory leaves open a free parameter that needs to be fixed based on the evidence, and the third to cases where a new theory is developed using the empirical results of an older theory. By exploring these types of cases, Worrall sought to develop an account of confirmation that gave prominence to use-novelty but nonetheless could be applied to actual scientific evaluations.

In later publications, Worrall abandoned the contingent use-novelty criterion (see, in particular, Worrall 2006, 2009, 2014). He argued that intuitions about scientific confirmation are actually best captured by a confirmation theory that is at root a logical one. The underlying sense in which predictions can be considered from a logical point of view can be illustrated usefully by how the word 'prediction' is used in science. Scientists sometimes use the word 'prediction' with reference to prediction of the future (i.e. 'forecasting'), but often they speak of prediction simply in a deductive sense where 'prediction' means that some empirical result "falls out" of the theory, rather than that result having had to be worked into the theory "by hand" (see Worrall 2009). For example, general relativity is said to 'predict' the perihelion of Mercury data, in that this data follows from the postulates of that theory (see Brush 1989). To give another, different type of example, psychologists might say that "self-esteem predicts happiness," meaning that there is a predictive statistical relationship between the variables 'self-esteem' and 'happiness.' From the logical point of view, 'prediction' picks out a certain way in which *the theory* stands in relation to empirical results, rather than whether some evidence was used by the theorist in constructing the theory.

To build a confirmation theory that is based on the logical sense of prediction, Worrall introduces the logical or structural notion of *a free parameter*, i.e. a theoretical constituent whose content is not tied down by the overall theory itself, but must instead be fixed into the theory based on empirical results as such (cf. Worrall 1985). 'Accommodation' is redefined using this notion: a theory logically accommodates evidence e when it contains

a free parameter that needs to be fixed based on e itself. 'Prediction' is simply the opposite of logical accommodation. A theory logically predicts evidence e when e follows deductively from the theory (and appropriate auxiliaries) but e has not been incorporated into the theory through parameter-fixing (see Worrall 2014, p. 55). Prediction and accommodation thus become a matter that concerns the logical or structural relationship between the theory and the evidence. To identify whether some evidence was predicted or accommodated, all we need to do is to examine the theory and the evidence as such: has the evidence been incorporated into the theory via a free parameter?

With the new logical definitions for prediction and accommodation, Worrall develops a Lakatos-inspired confirmation theory where confirmation depends on a three-place relation between a general theory T, a specific version of the theory T', and evidence e. These are related as follows. Consider a theory T that postulates some theoretical laws, mechanisms, or entities. The theory as such entails some empirical consequences (with auxiliary assumptions), but it can also leave open some empirical results. In these cases, the theory contains free parameters that need to be fixed based on the empirical results directly. For example, the wave theory of light (T) leaves open the wavelength of light from any particular monochromatic source. The theory does, however, allow for the wavelength to be experimentally measured from any particular source. In this way, using evidence e from such experiments, we can fill this parameter, and arrive at a more specific version of the wave theory (T'), where the correct wavelength is fixed directly into the theory.

This same procedure applies (or can be applied) to just about any theory or theoretical framework. For another, less scientific example, Worrall gives the 'Gosse dodge,' where Philip Gosse showed how the fossil evidence could be accommodated into young earth Creationism (T). First, we hold that the Earth was created very recently (generating the general theory T). Then, upon the discovery of fossils, we adjust our theory to hold that the Earth was created with the fossils already in place, which only gives it the appearance of having been created long ago. Here, the general theory (T) does not entail anything about the fossils as such. However, T can be amended to yield a more specific version of the theory T', where the fossils are added in via free parameters (i.e. young earth Creationism + fossils).

Worrall observes that we appear to be faced with a puzzle about scientific confirmation. On the one hand, accommodation through parameter-fixing is a perfectly legitimate scientific practice (e.g. the wave theory of light). Scientific theories often leave open free parameters that are fixed based on empirical observations. On the other hand, it is clear that any theory can be made to accommodate any evidence simply by adding more and more free parameters (e.g. the Gosse dodge). In other words, in so far as parameter-fixing counts for genuine epistemic support, any theory can acquire it with ease – scientific or unscientific.

Worrall's solution is to appeal to the logical distinction between prediction and accommodation. He argues that accommodation by parameter-fixing does indeed confirm something: *the specific theory T'* that has been obtained through parameter-fixing. But, it confirms T' only on the condition that there is epistemic support for T itself. In other words, parameter-fixing provides *conditional* epistemic support for a particular version T' of some general theory T, while T itself gains zero support from the accommodated evidence. *Un*conditional epistemic support for T itself, argues Worrall, requires logical prediction (i.e. the opposite of logical accommodation). This explains the epistemic asymmetry between (logical) prediction and accommodation, and reveals the relevant, underlying issue behind the predictivist intuition.

Worrall provides two more specifications for when logical prediction generates unconditional support for the general theory. In the first type of case, the theory is indeed amended by adding free parameters, but the theory with its fixed parameters proves to be independently testable. In other words, the accommodating assumptions turn out to generate further testable predictions. An example favored by Worrall is the prediction of the existence of Neptune based on Newton's theory. The Newtonian system, along with an auxiliary assumption about the number of planets in the Solar System, originally entailed the wrong result about the orbit of Uranus. Adams and Leverrier amended the auxiliary assumption, hypothesizing the existence of one more planet: Neptune. When the existence of Neptune was confirmed, Worrall argues that this did not just provide confirmation to a specific version of Newton's theory, but to the general theory itself.

The other type of case is defined more loosely. Worrall argues that theories also gain unconditional support in cases where empirical consequences "drop naturally out" of the theory (see Worrall 2014, p. 57). This happens, roughly, when the empirical results follow directly from some basic assumptions of the theory (along with some 'natural' auxiliary assumptions). Worrall's standard example is the Copernican theory and its explanation of planetary stations and retrogressions. The planets of the Solar System have been known for centuries to have additional motion in relation to the stars. This motion is usually eastward, but occasionally a planet slows to a halt, and then appears to move westward for a short period of time until it resumes its eastward motion. These anomalous motions were famously incorporated into the Ptolemaic system through the addition of free parameters, the epicycles. However, the motions of the planets are a natural consequence of the Copernican theory: we are on a moving observatory, the Earth, which occasionally overtakes slower planets or is overtaken by faster planets. The apparent slowing down, stopping, and retrogression of the planets is a simple consequence of the fact that the Earth itself is moving among the planets. For Worrall, the fact that the Copernican theory entails these empirical results in such a natural way provides *un*conditional support for it.

Apart from the conditional support that is obtained through parameter-fixing and unconditional support that comes through logical prediction, Worrall's theory leaves no room for further confirmatory distinctions based on prediction and accommodation. In other words, factors such as temporal novelty, theoretical novelty, and (contingent) use-novelty are not an issue in Worrall's account. Worrall argues that his logical account covers what is actually important behind these criteria. However, it is useful to recognize that Worrall's account does produce confirmatory verdicts that *correlate* strongly with accounts that place prominence on the distinction between (contingent) use-novelty and accommodation. Any evidence that follows from a theory but was not used by the theorist in its construction meets, by definition, Worrall's criterion for logical prediction, as this evidence will not have been used for parameter-fixing. (Similarly, any temporally novel evidence will automatically meet Worrall's criterion.) Given this connection, it is perhaps not surprising that Worrall's account is still often referred to as a use-novelty account, where he has redefined what use-novelty means (see Worrall 2009; Steele & Werndl 2018). To highlight the distinction to the standard use-novelty definition, I will refer to Worrall's overall theory as a 'parameter-fixing account,' and identify his version of use-novel prediction by adding the term 'logical' in what follows.

Worrall's account has inspired a lot of philosophical commentary. One of the most significant disputes over Worrall's theory concerns the competing theory of Mayo (1996). Mayo, as noted before, disagrees with Worrall about the underlying issue behind the prediction versus accommodation problem. She argues that whether or not rules about the use of evidence are legitimate depends on the severity of the test that they pose to the theory (see section 2.3). Worrall (2009) responds to Mayo, making the case that the underlying factor of importance is in fact his logical use-novelty criterion rather than test severity. Worrall argues that in all purported cases where this criterion appears to be violated (e.g. the previous SAT-example, or Howson's ticket example), there is actually a set of certain very general principles that are legitimately taken for granted (e.g. analytic principles), and appealing to these we arrive at a specific version of the theory (T') based on the data. Thus, these are cases that are covered by Worrall's logical theory of confirmation after all.

A number of commentators have considered Worrall's response unconvincing on this particular point (see Mayo 2009; Schurz 2014, p. 89; Votsis 2014, p. 74) – a reaction that I am inclined to agree with. An example such as inferring the average SAT-score of students in the class, even if artificial, appears to be a clear case where a particular hypothesis – i.e. 'the average SAT score of students in this class = x' – is indeed constructed from, and also confirmed directly by, the data. It is of course true, as recognized by both Mayo (2009) and Worrall (2009), that no theory or hypothesis is ever confirmed directly without the need for *any* auxiliary assumptions. But, it is difficult to see why the rules of arithmetic should be included here in what are

taken as parts of the theory that is fixed into a more specific version based on the data about the students, rather than simply among the set of (very general) auxiliary assumptions that apply in these types of cases. The SAT-example seems to successfully raise the concern that Worrall's theory may break down in certain cases.

Mayo's purpose in introducing the SAT-example is, still, only to elicit the intuition that using data to both construct and confirm hypotheses can be perfectly acceptable (see Mayo 2009, pp. 163-164). Many more similar examples can be introduced from scientific practice. Mayo highlights examples from statistics (see Mayo 2014), and more have been uncovered by Steele and Werndl (2016, 2018) (see below). To provide one recent example, consider the following one, adapted from Hollenbeck & Wright (2017, pp. 6-7). A team of epidemiologists test a new treatment for a novel life-threatening disease. They recruit a sample group of patients, half of which act as the control and half of which get the novel treatment. After the study, the results show a small, statistically insignificant average treatment effect between the control and the treatment group, i.e. the evidence fails to overturn the null hypothesis. However, a closer inspection of the data shows that there is a particularly strong, statistically significant effect among women in the treatment group. Furthermore, the effect becomes stronger the higher the woman's estrogen level. Background knowledge does not indicate anything particular for or against the hypothesis that the novel treatment interacts with estrogen to provide a cure to the novel disease. But, given the compelling data, the epidemiologists select this as their new hypothesis. Here, the only basis to construct the new hypothesis is the sample data, and indeed the sample data constitutes the only active reason to believe the hypothesis (apart, again, from auxiliary assumptions such as rules about inductive inference in science). I submit that here we have a case where the core content of the theory is fixed and simultaneously supported by the evidence.[10]

Another way to frame the issue of whether violations of the logical use-novelty rule could be acceptable proceeds in terms of the so-called 'no double counting rule' (see Mayo 2014). Worrall's position is often interpreted to entail a certain prohibition on the double use of evidence: you cannot use the same evidence twice, once in the construction of the theory and then again to support it (see Worrall 2009; Mayo 2014). In terms of Worrall's parameter-

[10] Worrall, I take it, would object that this is a case of 'deduction from the phenomena,' where what I take as the general theory (T) is inferred from some very basic principles – in this case, rules of statistical inference. Accordingly, the theory should be seen, after all, only as a specific theory (T') (see Worrall 2009). Here, I can only restate that I fail to see why rules about statistical inference should be included as content-parts of the actual hypothesis or theory that is being tested by the evidence. It is worth noting that Worrall himself recognizes a distinction between natural auxiliaries, the general theory, and evidence that unconditionally confirms the general theory in his examples of prediction of the second type where he discusses cases in which consequences "drop naturally out" of the theory (see Worrall 2014, p. 57). Could the example that I have given thus correspond to *these* type of cases after all? Here, we arrive at the difficult notion of *naturalness*, which Worrall has not attempted to clarify (see Worrall 2014, p. 57, fn. 5). The difficulty in specifying this notion generates further problems to Worrall's account, which I will return to in Chapter 6.

fixing account, this means that evidence that is used to fix free parameters can never unconditionally confirm the general theory. In response, Steele and Werndl (2016, 2018) argued recently that use-novelty and the no double counting rule do not always coincide: these can come apart depending on the overall theory of confirmation one favors as well as the type of scientific issue at hand. For example, in the statistical practice of cross-validation, a certain kind of use-novelty is important but there is also double counting of evidence. Steele and Werndl argue that no prominent theory of confirmation fully endorses Worrall's claim that only (logically) use-novel evidence can confirm the overall theory: there are legitimate cases of double counting in science (see Chapter 4 for further discussion).

Schurz (2014), seeking to improve on Worrall's account, develops another objection to Worrall's idea that there is a distinction between two different types of confirmation. Schurz points out that there are actually different types of parameter-fixing cases in science. In some cases, even though the overall theory does not settle a parameter value directly, it nonetheless constrains its contents in some way. For example, in the case of the wave theory of light, the overall theory already entails certain results about the experimental setting in which the free parameter 'wavelength' is measured. The fact that the experiments yield results that fall within these constraints, Schurz argues, should count for some *un*conditional support for the general theory (see also Schindler 2014, p. 65). He argues that the actual problematic case in which there should be no confirmation of the general theory is when the theory could have been fitted to *every possible data* through parameter-adjustment.

Schurz proceeds to develop a novel account of 'genuine confirmation' without Worrall's conditional/unconditional distinction which incorporates the insight about different types of parameter-fixing cases. Schurz introduces some logical machinery to distinguish between the content-elements C of theories T, and argues that in cases where evidence e increases the probability of T, the confirmation spreads to any C_i if C_i is *necessary* within T to make e highly probable. Thus, parameter-fixing cases where the theory constrains the contents of the parameters in some way can count for overall confirmation of important content parts of the theory. If, in contrast, T is amended with a parameter-adjustment of C_i to e that would have allowed for any empirical result, no confirmation of T occurs. Schurz argues that another improvement that can be made to Worrall's account is to apply it also to inductive, probabilistic cases – namely, curve fitting (cf. Hitchcock & Sober 2004). In these cases, to confirm a curve that has been fitted to a dataset E_1, Schurz holds that it must be tested with another use-novel dataset E_2. Given that the role of E_1 and E_2 can be reversed, Schurz holds that a full account of genuine confirmation must also make room for propositions describing the procedural role of evidences. Schurz's proposal could therefore be seen as a compromise between the logical and methodological approaches to confirmation (and predictivism), where logical considerations form the core of

the confirmation theory, but there is space for a contingent, methodological component.

Further criticism of Worrall's approach has been leveled both from the perspective that it is *too* contingent as well as that it is not contingent enough. Votsis (2014, pp. 74-75) argues that the fact that in Worrall's account it is possible for the confirmatory role of two datasets to be reversed (as shown above) means Worrall has failed to deliver a fully logical theory of confirmation, which Votsis considers an important cornerstone of an objective theory of confirmation. Pointing to a similar example as above, Votsis argues that it is possible for there to be cases where we have two qualitatively indistinguishable datasets, O_1 and O_2, each of which is sufficient to fix free parameters in theory T. If the free parameters are fixed based on O_1, and T' is then used to predict O_2, T gets conditional support from O_1 but unconditional support from O_2. However, if the role of the datasets had been reversed, T would have gained unconditional support from O_1 instead. Votsis argues that Worrall's theory has failed to meet the standard Worrall himself has set: "same evidence, same theory, same confirmation" (see Worrall 2006, p. 55).[11]

Douglas and Magnus (2013, pp. 587-588) make the exact opposite case to Votsis. They argue that in Worrall's account there is not enough room to capture the various merits of contingent novel prediction. They argue that Worrall's theory requires a clear mathematical structure to the theory and some way of arriving at a clear assessment of its free parameters. They hold that this clarity is often lacking in practice, and for this reason novelty, particularly temporal novelty, can provide further assurances that the theory is on the right track. In other words, Douglas and Magnus argue that despite his efforts of clarifying the epistemically relevant distinction between prediction and accommodation, Worrall's account still does not enable uncontroversial assessments of (logical) use-novelty, and for this reason, temporal novelty can provide stronger confirmation.

Worrall's parameter-fixing account is the only account in the literature that has been developed explicitly as a logical predictivist theory. However, in this context, one other theory bears discussing; namely, that of Lange (2001). Lange argues for a predictivist advantage by appealing to the notion of 'arbitrary conjunction.' Lange observes that when a hypothesis is judged to consist of an arbitrary conjunction of unrelated propositions, the hypothesis is typically not considered to gain much support from the pieces of evidence it entails. For example, the conjunction 'gravity bends light and the moon is 385 000 kilometers away from Earth' is an arbitrary conjunction, where providing evidence for either part of the conjunction does not tell us anything about the other part. Lange argues that the problem with theories

[11] Worrall (2006, pp. 51-56) considers objections of this type, which Votsis (2014, p. 75) discusses but finds Worrall's responses wanting. Worrall's argument is basically that in these case it is shown that there is one 'unit' of genuine unconditional confirmation for the theory. Votsis (ibid.) replies that this is "tantamount to burying one's head in the sand," as there is still the problem that O_1 and O_2 can yield conflicting confirmatory judgments depending on their contingent role.

that have been constructed through accommodation is that they have a tendency of becoming arbitrary conjunctions. That is, when accommodations are made to theories, they tend to become more disjointed as a result. Novel predictive success provides evidence that a hypothesis is *not* an arbitrary conjunction, and in this way novel predictions can provide stronger confirmation than accommodations in science.

Interestingly, Lange intends for his theory to be interpreted *ontologically*. He holds that what gets confirmed in cases of prediction is the theory rather than facts about what the theorist may or may not have done in constructing the theory (see Lange 2001, p. 582). In other words, Lange appears to be going for a logical theory of confirmation where what matters is the logical or structural relation in which the evidence stands to the theory. However, there are some problems in interpreting his position. Lange does not take a stance on how scientists make their judgments about whether a hypothesis constitutes an arbitrary conjunction or not. On the one hand, his theory thus leaves open the possibility that this judgment cannot always be made reliably, which would mean this theory requires a supplement where information about how a hypothesis was constructed plays an epistemic role (see, for example, Lipton's 'fudging' theory). On the other hand, if Lange holds that judgments about arbitrary conjunctions *can* be made reliably in science, his theory removes the epistemic role of prediction and accommodation in themselves. What matters is simply whether a hypothesis is an arbitrary conjunction, no matter how it was constructed.[12]

Worrall (2014, pp. 60-61) has recently challenged Lange, arguing that the notion of arbitrary conjunction, as an underlying epistemic factor in science, is untenable. Worrall points out that any pattern of data that seems arbitrary can be considered non-arbitrary in infinitely many ways, and whether or not it is considered non-arbitrary will always depend on *theory*. For example, a dataset that consists of recordings of the position of Mars may seem entirely arbitrary, until we have Kepler's laws (or later improvements) that show the underlying pattern behind the recordings. In a similar way, all kinds of patterns can be argued to be non-arbitrary as far as enough parameter adjustments are made. Worrall argues that Lange's theory is unable to capture a confirmatory distinction between cases where a theory achieves a fit with the evidence *without* parameter adjustments and cases where the theory requires such adjustments – the cornerstone of Worrall's own theory.

Worrall's objection, in my view, is compelling. The work that Lange's notion of an arbitrary conjunction is meant to do is captured more satisfyingly by Worrall's parameter-fixing account, which has the resources to distinguish between appropriate and inappropriate cases where evidence is tacked into a theory in more or less artificial ways. Furthermore, if Lange holds

[12] A fully logical version of Lange's theory is a version of what Barnes (2008) has called 'thin predictivism.' According to this interpretation, Lange holds that accommodation *correlates* with arbitrary conjunctions, but accommodation and prediction do not have an epistemic role in themselves, as whether a hypothesis constitutes an arbitrary conjunction or not is evaluated directly.

that arbitrary conjunctions cannot be evaluated as such in science, the lesson the theory provides about the predictivist advantage appears to be already covered by Lipton's (2004) fudging account. If, in contrast, Lange argues that arbitrary conjunctions can be evaluated as such, prediction and accommodation no longer have an epistemic role in themselves. For these reasons, I will concentrate on Worrall's account as the most promising logical predictivist theory in this study.

To conclude on the logical approach, this approach has been introduced and developed chiefly by Worrall (2006, 2009, 2014). A number of commentators have considered Worrall's theory promising on the issue of predictivism (see Mayo 2014; Schurz 2014; Votsis 2014; Schindler 2014; Steele & Werndl 2018). However, several problems have also been raised, three of which appear the most pertinent. First, the distinction Worrall introduces between conditional and unconditional confirmation may break down in certain common cases (e.g. Mayo 1996; Steele & Werndl 2016, 2018): there are cases in science where evidence that is directly needed to fix the contents of the theory appears to confirm that theory, contrary to Worrall's parameter-fixing account. Second, questions can be raised whether the distinction between conditional and unconditional confirmation is needed to start with. There can be degrees in the extent to which theories constrain the contents of free parameters, and we may only need one type of confirmation to capture this (see Schurz 2014). Third, we can ask, is Worrall's account able to cover all that is relevant about prediction to scientific confirmation (see Douglas and Magnus 2013, pp. 587-588)? The previous section introduced multiple methodological issues with the use of evidence that are not included in Worrall's account (and more potential advantages to prediction are discussed below). Are the competing predictivist theories correct in that there are further epistemically relevant aspects to scientific prediction than those covered by Worrall's parameter-fixing account? We will return to reconsider Worrall's theory and the logical approach in Chapter 6.

2.5 THE AGENT-BASED APPROACH

The final type of weak predictivist approach is the agent-based approach. This approach is characterized by what has been called 'the epistemic rebound effect,' where novel prediction provides evidence for the theory *via* the evidence it provides about the theorist's reliability. The agent-based theorists take to heart the difficulty in justifying strong predictivism through an explanatory argument (see section 2.1). The best explanation for *novel* success, it appears, cannot just be the truth of the theory, because the truth of the theory in itself does not have any bearing on novelty, which is a contingent issue that concerns the theorist's use of evidence (see Barnes 2002a; White 2003). So, the agent-based theorists argue that in cases of novel predictive success, rather than inferring directly that the theory is likely to be true,

evaluators should infer that the theorist possessed some type of epistemic virtues, goods, or properties that made him more likely to be a reliable source of true theories. The fact that the theorist selected or generated the theory in a reliable way *can* explain novel success, i.e. how the theorist was able to find a theory that entails some evidence that was not used in the construction of the theory. And, when there is evidence that the theorist is likely reliable, it can *then* be inferred that the theory is more likely true. With accommodative success, there is no need for a similar reliability-based explanation. Accommodative success can (presumably) be explained just by pointing to the theorist's intention of accommodating the evidence, so no further advantages (or disadvantages) arise for theory confirmation.

The agent-based approach divides further into two subcategories. The most recent agent-based theory of Barnes (2008) is developed around a novel definition of novelty, which differs significantly from the standard use-novelty conception. According to Barnes, what matters in the predictivist argument is the theorist's *novel endorsement* of the theory (and its predictions) rather than the use-novelty of the evidence. There are a number of important differences between the use- and endorsement-based agent theories, so a separate discussion of each is warranted here. To signify the difference between use- and endorsement-based conceptions of novelty, the former type of agent theories are called UN-theories and the latter an EN-theory in what follows. I will start with the UN-theories.

UN-theories. A UN-based predictivist theory has been developed both within Bayesian and IBE-based confirmation theory. Maher (1988, 1990, 1993), who presents the Bayesian version of the argument, was the first philosopher to argue for an agent-based theory on the prediction versus accommodation issue. Maher's predictivist theory is based on two insights. First, he proposes that in the scientific process, theorists use some type of 'discovery methods' in coming up with their theories and hypotheses, and these discovery methods can be more or less reliable.[13] In other words, Maher turns attention from the theory to the theory generation process, and its reliability in generating (true) theories. Second, Maher argues that to gain evidence about the reliability of a theorist's discovery method, it is possible to refer to novel predictive success: novel success indicates that the theorist used a reliable discovery method in generating their theory. Appealing to these insights, Maher (1990) provides a Bayesian proof for the predictivist intuition. In this proof, the theorist's degree of reliability is interpreted as his or her probability of success in generating a successful theory, a probability which is expected to lie somewhere between absolute certainty and impossibility.[14] The

[13] Maher uses the word 'method' in his papers, but this should not be confused with what has here been called the methodological approach to weak predictivism. Maher is talking about *whatever* it is that enables a *particular scientist* become a reliable source of theories, not what the use of particular methods or procedures may do to cause the theory construction, selection, or testing process to become more unreliable in the case of any scientist.

[14] As an example of further assumptions of Maher (1990), he requires that the reliability of the theorist's discovery method is previously unknown to us. Maher (1993) responds to critique by Howson

proof shows that, under certain further assumptions, novel predictive success should drive up the expected value of the theorist's probability of success in theorizing, which in turn should drive up evaluators' posterior probability for the theory.[15]

White (2003) presents a similar, IBE-based theory. White formulates his theory in terms of an 'archer analogy,' where we are asked to imagine a theorist selecting a datum-entailing theory from a space of logical possibilities. If the theorist succeeds in selecting such a theory, a different explanation appears to be called for in the case of prediction than in the case of accommodation. In the case of accommodation, we know that the theorist was going for a datum-entailing theory all along, so no further explanation is required (cf. Worrall 1985). The fact that the theorist was going for such a theory explains why he now holds such a theory. With predictive success, on the other hand, some other explanation must be produced – for how could the theorist have found a theory that entails a particular datum when that datum was not known? White argues that the best explanation is that the theorist was *reliably aiming at the true theory* among the theories that entail the predicted datum. For if the theorist was reliably aiming at the truth, it would not be a mystery how he managed to pick a datum-entailing theory from all the possibilities. The evidence about the theorist's reliability is then used to infer that the theory is more likely true.

Maher's and White's theories are based on the same underlying idea that turns on how the theorist generates or selects her theory. When the theorist does not use evidence in generating an empirically successful theory (i.e. the theorist predicts evidence instead of accommodating it), Maher and White argue that this success calls for an explanation in terms of the theorist's reliability. In other words, in these theories, *the evaluator's belief in the theorist's degree of reliability* becomes the factor that delivers the predictivist advantage. A sufficiently high degree of reliability will explain the novel predictive success, and enable evaluators to infer that the theory is (more) likely true as well. The agent-based predictivist boost is then added to the epistemic support for the theory along with other sources of support (e.g. the evidence as such), the existence of which Maher and White are naturally not denying (see Maher 1990, p. 328; White 2003, pp. 671-674). The idea is, basically, the following. Suppose that we have a case where the evidence itself justifies a probability of 0.5 for the theory. Next, we learn that the theorist has achieved novel predictive success. Due to this success, there is now a reason to

& Franklin (1991), adding the further condition that the output of the method must also be previously uncertain. Previously, Maher (1988) provided a proof of concept, making the highly idealized assumption that the theorist's discovery method must be either completely reliable or completely random.

[15] Kahn, Landsburg & Stockman (1992) have proposed a similar theory to Maher, starting from the same motivations. However, given that KLS (1992) employ a similar unrealistic assumption as Maher (1988) about there being only two types of scientists, those who are 'talented' and those who are not, I will refer to Maher's (1990) theory as the most advanced version of this type of Bayesian UN-theory. See also Barnes (2008, p. 120) on the talent assumption.

believe that the theorist's degree of reliability in theorizing is relatively high(er), say 0.8. This new information now enables us to bump up our confidence for the theory.

The UN-based approach to weak predictivism faces some difficult objections. First, this approach does not appear to accord with scientific practice, which has led some to dismiss it immediately. Lipton (2004, p. 167) argues that in science we are meant to evaluate theories rather than theorists, so a confirmation theory that refers to the theorist is a non-starter. Lipton's objection, even if brief, is not entirely without bite. For example, as far as I know, no textbook on scientific methodology encourages the working scientists to evaluate predictions by evaluating the degree of reliability of the theorist behind the theory: what seems to count are the theoretical assumptions, methodological choices, and the evidence presented.[16] (To be sure, there may be reasons to be suspicious of the work of particular scientists. However, in these cases, it seems that the predictions achieved by these scientists could themselves be suspect rather than confirmatory of reliability.[17]) In other words, the kind of epistemic rebound effect that the agent-based theorists postulate does not appear to be a factor in scientific practice, or at least not a lot of evidence has been presented that it is.

Second, the UN-based approach is quite vague, and it is difficult to see if it could be made precise enough to support actual confirmatory evaluations in scientific practice. The UN-based agent-theorists propose that in a case of novel predictive success, this success implicates some particular degree (or range) of the reliability of the theorist that then reflects favorably on the theory. The challenge is, how can this degree of reliability be determined and applied to confirmatory evaluations in any clear or uncontroversial way. Howson and Franklin (1991) argue against Maher that it cannot: the very idea that there is a well-defined chance of an evidential statement (i.e. a prediction) and a theory being true conditional on having been generated by a particular attempt at theorizing stretches, according to them, the notion of chance to "quite impossible lengths" (ibid. p. 583). However, even if it could be established that there is a well-defined degree of the reliability of the theorist that is implicated in particular cases of novel success, there is still the problem of how this degree of reliability can be evaluated in any dependable way. Worrall (2014, p. 58, fn. 6) argues that in so far as the discovery method that the theorist uses in deriving their theory cannot be rationally reconstructed – which has not been attempted by Maher and White (although see Barnes 2008, pp. 82-121) – there do not appear to be any clear confirmatory judgments to be made about it (for example, how do

[16] This point may be vulnerable to what Barnes (2008, p. 83) has called "the classic fallacy of social science" in reference to a similar argument about the prediction versus accommodation issue: the view could be based on what members of a community say they are doing rather than what they are actually doing. Evidence from the sciences is evaluated in more detail in Chapter 4.

[17] For one such example, see Brush (1995, pp. 306-307). Brush writes that the research program of the physicist Hannes Alfven was rejected *despite* his novel predictive successes, on the grounds that Alfven's basic assumptions were considered unacceptable to other physicists.

we justify believing that the theorist's degree of reliability in a particular case is 0.8 rather than, say, 0.6, or 0.2?).

The final issue with the UN-agent-based approach is that the problems of the traditional strong predictivist argument notwithstanding, it still appears strange and contrary to reason to credit novel predictive success to the theorist rather than directly to the theory. Worrall (2014, p. 59, fn. 7) discusses a case of earthquake prediction, where a theorist successfully predicts nine earthquakes in a row. He points out that the novel predictive success in this case can obviously only be achieved if the theorist is working with a well-confirmed *theory* about seismic activity. The idea that we should think of the theory as confirmed by the predictions only *via* inferring that the theorist must have generated the theory in a reliable way appears a needless detour. To put the intuition another way, even if we believed that the theorist generated the theory in some *un*reliable way (say, he or she made some shaky assumptions along the way to deriving it), it still appears that the consistent predictive success confirms that the theory is indeed sound, no matter how it was originally derived. (This is a result that is entailed by the confirmation theories of Worrall and Mayo, for example.) In other words, *intuitively*, the predictive success appears to confirm the theory, which forms the basis of those predictions, regardless of the reliability of the way in which the theorist originally arrived at it.

The UN-based agent-theorists are not without options in how to respond to these problems. First, the approach is based on an explanatory argument, which has not been challenged directly. The UN-based agent-theorists argue that the theorist's reliability is *the best explanation* for the achievement of novel predictive success. Nothing that has been said so far yet challenges this claim by providing a better explanation: how do novelly successful theories and ideas emerge in science if not via a reliable route? Second, the agent-based approach can also be subjected to more detailed evidence about scientific practice. Whether theorists are indeed appropriately reliable in cases of novel success can, at least to some extent, be evaluated against the historical record of scientific theorizing. Perhaps, closer attention to theory generation in science could reveal that novelly successful theorists generally *are* reliable in some clearly evaluable way. For these reasons, we will briefly revisit the UN-agent-based approach in Chapter 5, where a more detailed assessment is conducted of the overall explanatory strategy in defending predictivism (and scientific realism).

Barnes's EN-theory. Barnes's (2008) EN-theory is based on a similar intuition about the role of the theorist in the predictivist argument as the theories of Maher and White. However, Barnes introduces a number of new ideas into the agent-based approach. First, Barnes criticizes Maher and White for their reliance on what he considers vague new ideas about scientific discovery methods and argues that a better, more precise approach is to go with something that is more in line with standard confirmation theory: *(true) background beliefs* (see Barnes 2008, pp. 115-116, 120-121). For Barnes, what

makes theorists reliable or truth-conducive in theorizing is not a discovery method but a set of true background beliefs. Second, Barnes argues that it is not actually the process of constructing the theory that is of epistemic importance but the fact that the predictor *endorses* the theory.[18] By endorsing the theory, the theorist provides a signal to evaluators that he has some good reasons (i.e. a set of true background beliefs) to think that the theory is true. Finally, Barnes argues for something he calls 'epistemic pluralism.' Epistemic pluralism holds when subjects also consult other people's judgments about what to believe rather than just their own. Barnes makes the case that there are good reasons for scientists to be epistemic pluralists in the context of scientific practice. For instance, scientists may not always be aware of all the relevant evidence on a given subject, or they might be uncertain about some evidence and what it tells about a particular theory. This means there are good reasons for scientists to be interested in each other's judgments – and indeed the background beliefs that underlay those judgments.

Drawing on these insights, Barnes presents a distinctive agent-based theory for epistemic pluralists which he calls 'virtuous tempered predictivism.' In this theory, predictivism holds if three conditions are met: a) there is *a pluralist evaluator* who places epistemic importance on the judgment of others, b) there is *an endorser* who endorses a scientific theory more strongly than the evaluator (prior to the confirmation of some prediction), and c) the evaluator has *trust in the endorser's competence* in that the endorser would not endorse the theory unless he had good reasons to do so. The evaluator's trust in the endorser now becomes fundamental to the predictivist effect (see Barnes 2008, pp. 64-69). The predictivist advantage arises for pluralist evaluators in cases where an endorser posts a sufficiently high prior probability for a theory before its prediction (or predictions) is (/are) confirmed,[19] and the evaluator trusts that the endorser is 'competent' (i.e. the posted probability accords with what is justified by the endorser's actual evidence). Recognizing this predictive act, the evaluator raises his own prior probability for the theory, now approaching the endorser's probability. Barnes calls this *per se predictivism*. In addition to this, there is a further confirmatory effect which kicks in when the prediction is confirmed. At that point, even if the evaluator had some modest doubts about the endorser's background beliefs before, they are further reduced by the confirmation of the prediction. Barnes calls this *success predictivism*. He emphasizes that success

[18] Barnes's argument here is an intuitive one (see Barnes 2008, p. 35). He argues that in a case where we have novel predictive success but the theorist does not endorse the theory but instead holds that further predictions of the theory will fail, evaluators should believe what the theorist says rather than go by use-novel success. Harker (2011. p. 221) does not share this intuition, arguing that Barnes's example is not sufficient to show that other predictivists have been wrong in assigning importance to use-novelty instead of endorsement-novelty. He argues that in the type of case introduced by Barnes, the evaluators could, for example, reason that the theory might be veridical in some respects but not in others.

[19] Barnes (2008, pp. 35-36) argues that the prior probability must be higher than the evaluator's and either high in general "or at least not so low as to constitute a primarily skeptical or noncommittal attitude toward the theory."

predictivism holds only if *per se* predictivism holds: if the evaluator does not trust that the theorist has some good reasons to advocate the theory that entails the prediction, these reasons cannot subsequently be confirmed by the success of that prediction (ibid., p. 69).

Barnes's theory is based on a similar explanatory argument as White's and Maher's theories: the theorist's reliability or truth-conduciveness in advocating a particular theory, understood now through the possession of a set of true background beliefs (instead of the use of a 'discovery method'), is invoked by the evaluator as the best explanation for why the theorist advocates a theory that entails novel predictions. However, Barnes's transition from the theorist's use of evidence to the theorist's endorsement of the theory creates an important difference. Whereas for Maher and White what determines how novel prediction affects the probability assigned to the theory is the evaluator's belief in the theorist's degree of reliability, for Barnes this is the probability that the theorist posts for the theory.[20] In other words, in Barnes's theory, the evaluator updates his prior probability for the theory based on the theorist's probability instead of his belief in the theorist's degree of reliability in generating theories.

The transition from use-novelty to endorsement-novelty provides Barnes with a certain advantage over the UN-theories: he is now able to escape the criticism about the vagueness of the UN-approach. Whereas the theorist's degree of reliability in theorizing is difficult to evaluate, the theorist's own judgment about the probability that should be assigned to the theory is something that can, at least in principle, be expressed rather precisely and shared publicly. However, this very same move appears to cause an even deeper problem, which I am inclined to think is ultimately fatal to Barnes's account: as the focus shifts from construction to endorsement, the epistemic issue seems to shift from prediction and accommodation simply to evaluators' judgments about the trustworthiness of endorsers. This has been pointed out by Harker (2011), who argues that all that Barnes establishes is that, other things being equal, scientists should prefer theories that have been endorsed by reputable individuals.[21] I agree with Harker's conclusion, although I will present the argument in a different manner than he does. In my view, the conclusion follows directly in consequence of Barnes's transition from use-novelty to endorsement-novelty:

As we have seen, in Barnes's theory, the agent-based confirmatory effect arises based on the evaluator's trust in the endorser (see Barnes 2008, p. 69). When an evaluator believes that the endorser is trustworthy, and the endorser posts a higher probability for the theory than the evaluator, the

[20] Of course, as stated above, in Barnes's theory, the evaluator must also *first* believe in the theorist's trustworthiness, but when she does, the epistemic factor that affects theory confirmation for the evaluator is the probability posted by the theorist.

[21] Harker (2011) appeals to distinct priors that are assigned based on endorsement, which underlay the pertinent differences in Barnes's per se predictivism (and ultimately success predictivism). I believe that my argument is ultimately the same one that Harker makes, but I hope to make it clearer. Barnes (2014) responds to Harker (2011) by reiterating the pertinent features of his account.

evaluator is then able to raise their own probability for the theory. But, given that trust in the endorser is fundamental to the endorsement-based epistemic boost, *prediction and accommodation* are no longer the issue: in so far as the evaluator trusts the endorser, whether the endorser makes predictions or accommodations should make no difference. In the case of prediction, the evaluator is supposed to reason that the endorser must have some additional evidence for the theory; otherwise he would not post the high(er) probability for it. This is fair enough, and we may very well accept this. However, the same line of reasoning, it appears, should apply in the case of accommodation. If a trustworthy endorser posts a higher probability (i.e. higher probability than the evaluator) for a theory that is supported by accommodated rather than predicted evidence, the evaluator, given his trust in the endorser, should reason that the endorser must have some additional reasons to have confidence in the theory; otherwise he would not post the higher probability for it.[22] In other words, the confirmatory boost that arises from the evaluator's trust in the endorser is there in either case. In order to establish a difference between *prediction and accommodation*, some other argument is needed to make the case that a different weight should be assigned to the endorser's probability in the prediction case than in the accommodation case. And, this simply takes us back to the prediction versus accommodation issue. (A difference could be shown, for example, through the methodological approach. Evaluators might reason that given that the endorser has relied on accommodated evidence, he might have done something – either inadvertently or purposefully – that has caused the theory selection process to become more unreliable. Predictive success then alleviates these doubts and shows that the theorist's endorsement can indeed be trusted.[23] So, while the evaluator may put weight on the theorist's endorsement, the epistemic difference between prediction and accommodation is established based on the methodological argument.)

This problem appears to me fundamental: as far as endorsement, and trust in that endorsement, is what counts from the epistemic point of view,

[22] Barnes (2008) appears to seek to define endorsement-novelty in such a way that this is meant to be prevented: to make an endorsement-novel prediction of evidence e is not to appeal to e in endorsing the theory that fits e. To endorse a theory based on accommodation of e means that one does not have further true background beliefs that are sufficient for endorsement. However, the fact that an endorser can endorse a theory more strongly than the evaluator based on accommodation is a possibility that follows from Barnes's epistemic pluralism. For example, the endorser may need the accommodated evidence for actual endorsement, but he may also have more additional evidence than the evaluator that is combined with the accommodated evidence, or the endorser could be in a position to see the evidential impact of the accommodated evidence more clearly than the evaluator. In this, we see that the trustworthiness of the endorser is what counts rather than the prediction vs. accommodation distinction, as Harker (2011) argued.

[23] This same result is also entailed by Barnes's success predictivism, where the success of the predictions alleviates remaining doubts about the predictor. However, in this, Barnes's theory is already reliant on an existing predictivist argument rather than presenting a new one (i.e. the negative, methodological one). Altogether, I suggest that the predictivist aspect of Barnes's endorsement-based confirmation theory reduces to a less than optimally formulated negative predictivist account.

prediction and accommodation remain on a par until a predictivist argument is added to Barnes's endorsement-based theory. There are also some further problems. Chiefly, there is also an issue related to the idea that endorsement-novelty rather than use-novelty is epistemically central or important. Contrary to Barnes's assumptions, endorsement-novelty unlike use-novelty does not appear to be a fact that calls for any 'deep' explanation to begin with (e.g. in terms of reliability or truth-conduciveness). Endorsement is simply a decision that the individual makes, which may be affected by all kinds of contingent factors other than the evidence, such as personality traits, a desire to appear modest to one's peers, a desire to appear highly confident, etc. The UN-theories have a certain intuitive advantage in that the fact that a theorist has been able to generate a successful theory without relying on some evidence is an issue that genuinely appears to require some explanation: how *did* the theorist arrive at such a successful theory? In contrast, with endorsement-novelty, there can be any number of competing explanations for the endorsement but the truth-conduciveness of the theorist (e.g. 'overconfidence'). A similar point is made by Glymour (2008), who points out against Barnes that "the history of science is a history of endorsements on what seem, in retrospect, insufficient evidence."[24] Similarly, Leplin (2009) argues that endorsement prior to the actual testing of novel predictions may simply *discredit* the endorser rather than provide evidence of their trustworthiness.

To conclude, Barnes's EN-theory, despite having been developed as a predictivist account, does not establish a new argument for predictivism. I agree with Harker (2011), who argues that Barnes's theory is better seen as a theory about endorsements rather than prediction and accommodation. Barnes successfully raises the issue that rather than simply focusing on the theory and the evidence as such, scientists may also be interested in, and put weight on, each other's opinions about what the evidence implies about the theory. This is a view on confirmation that applies independently of the prediction versus accommodation issue, and Barnes (2008) develops multiple new arguments about how this insight could be utilized in philosophical theories about scientific confirmation.

2.6 OTHER CONTEMPORARY APPROACHES TO PREDICTIVISM

In addition to the vast literature on weak predictivism, there are a few contemporary accounts that extend beyond a defense of a particular weak predictivist theory.

[24] Barnes (2008) holds that 'competent' scientists do not endorse without sufficient evidence. However, as pointed out by Glymour (2008), this does not appear to match with what happens in scientific practice, where even highly competent scientists have very often been wrong. For some interesting examples of confident endorsements of the wrong position in science, see for example Frankel (2012a, pp. 118-119, 139-142).

Douglas and Magnus (2013) argue for what they call 'Pluralist Instrumental Predictivism.' According to this account, the value of novel prediction does not reduce to any one particular virtue, but to many different virtues at different levels of inference. Douglas and Magnus discuss inferences from data to phenomena, phenomena to theory, and theory to framework, and argue that in each case novel prediction provides advantages over accommodation. Here, data are understood as raw observations (e.g. points on a scatter plot) and phenomena as patterns in the world indicated by data, following Bogen and Woodward (1988). Theories explain or predict phenomena, and 'framework' is the overall background against which particular theories are generated. On the level of inference from data to phenomena, Douglas and Magnus argue that novel prediction provides assurances that the data has not been overfit. Their argument is otherwise similar to Hitchcock and Sober (2004), but they argue further that *temporally* novel prediction is better than accommodation with AIC, because it is often not clear whether the assumptions behind AIC hold in practice. On the level of inference from phenomena to theory, Douglas and Magnus argue that the ability of theories to predict novel phenomena provides some inductive support for the capacity of these theories to guide us to further such successes. Furthermore, novel success provides evidence that we have not 'fiddled around' too much with the core structure of the theory to make it fit known phenomena (cf. Lipton 2004). Finally, on the level of inference from theory to framework, the generation of novelly successful theories gives us a prima facie reason to adopt the framework based on which the theorist generated it (see Douglas & Magnus 2013, p. 585-586; cf. section 2.5).

Alai (2014, 2016) advocates a new, distinctive version of *strong predictivism* that is based on a novel 'functional' account of novelty. Alai explains that a theory makes a functionally novel prediction of datum d if a) d has not been used *essentially* in the construction of the theory (i.e. the theory and the auxiliaries are plausible independently of d), b) d is *a priori improbable*, and c) d is *heterogeneous* to the essentially used evidence. Roughly, a prediction is thus functionally novel if it is precise (in a logical sense) and concerns some different type of phenomenon than those that were needed in the construction of the theory. Functional novelty, Alai argues, respects the logical or deductivist view of confirmation in that the functional novelty of evidence does not depend on contingent factors such as who constructed the theory and for what reason (in contrast to use-novelty). Yet, functional novelty is also not a purely logical notion, as it depends on the overall epistemic situation in which we evaluate the relevant criteria for functional novelty. In this, functional novelty can also capture the important insight of the predictivists that not all consequences confirm the theory equally. Alai proceeds to argue that the functionally novel consequences of scientific theories warrant an inference to the truth of the theory via a particular chain of explanatory inferences. He holds that the best explanation for functionally novel success is that A) the theorist picked a theory that

contains enough truth to have the functionally novel evidence as its consequence. A, on the other hand, is explained by B) the theorist, using the scientific method, aimed at a true theory in a reliable way. And B is explained by C) the scientific method is reliable because it (i) assumes that nature is simple, uniform, and intelligible and hence knowable through analogy, induction, and abduction, and (ii) nature actually is simple, uniform, and intelligible (see Alai 2016, p. 554).

Douglas and Magnus combine multiple predictivist arguments to introduce a more holistic framework in which they argue for various advantages of novel prediction over accommodation. Alai combines multiple insights from different points of view on the prediction versus accommodation issue to develop a novel defense of strong predictivism and scientific realism. I will return to evaluate also these accounts at different points in the following chapters, among other predictivist arguments and approaches (see, in particular, Chapters 4 and 6).

2.7 ANTI-PREDICTIVISM

As opposed to the myriad of arguments that have been developed in support of predictivism, there are also challenges to it. These challenges can be categorized roughly into two types: (i) *principled* and (ii) *generalist*. The former type of challenge rejects predictivism based on an overall principle or argument. Basically, this strategy amounts to arguing that contingent factors should not be included in confirmation theory. The second type of challenge approaches the issue from a broader perspective. The gist of the more general challenge is to argue that novel prediction is not anything special. There can be many sources of epistemic support for a scientific theory, among which novel prediction does not stand out in any particular way. This approach does not necessarily reject predictivism in all circumstances; it simply argues that on the balance, there is no strong reason to prefer novel prediction to accommodation. In fact, accommodation could even be better all things considered.

The principled approach. The principled objection to predictivism has a long history in the philosophy of science. It is rooted in the logical empiricist program and the goal to develop a purely logical theory of confirmation (see Musgrave 1974, p. 2). From the logical point of view, scientific confirmation should be viewed as an *objective* matter that concerns the relationship between the theory and the evidence. According to this view, theories gain epistemic support from their true empirical consequences. If there are false consequences, these count for disconfirmation. Whether the theory or the evidence came first is completely irrelevant (cf. Keynes 1921).

In contemporary confirmation theory, this position is typically associated with a strict Bayesian approach to confirmation (see, for example, Mayo 1996, pp. 319-360; Howson & Urbach 1996). According to the Bayesian

approach, our degrees of belief are represented by probabilities, which are updated based on the Bayesian formula, in a process called *Bayesian conditionalization*:

$$Bayesian\ formula\colon P(H|E) = P(E|H) * P(H)/P(E)$$

In the Bayesian formula, P(H) represents our *prior* belief in the hypothesis H, and P(H|E) is our *posterior* belief in H that is updated based on the evidence, E. If P(H|E) > P(H), E (incrementally) confirms H. The strict Bayesian holds that scientific confirmation is captured fully by the Bayesian calculus, where prior probabilities are updated based on the evidence using the Bayesian formula (or, some version of it). There is nothing in the formula itself that reserves a role for the novelty of evidence, so the novelty condition becomes irrelevant to confirmation.

The strict Bayesian position enables a simple argument against predictivism: according to the Bayesian calculus, using evidence to both construct and confirm theories is perfectly acceptable. In so far as Bayesianism provides a compelling overall approach to confirmation, so much the worse for predictivism.[25]

There are at least two major problems in appealing to Bayesianism to refute the predictivist position. First, as noted before, the Bayesian framework is malleable enough to accommodate any view on the predictivism issue (see Brush 1994, p. 134). Bayesian confirmation theories come in many different forms, ranging from subjective (e.g. Howson & Urbach 1996) to objective (e.g. Williamson 2010), and whether or not predictivist intuitions are captured in the Bayesian calculus can simply depend on how one interprets the Bayesian formula.[26] There is also nothing in principle about the Bayesian approach that forbids that facts about the use of evidence are *added* into the Bayesian calculus (see, for example, Maher 1990; Schurz 2014, p. 91). So, the Bayesian formula does not settle in itself the question of whether prediction or accommodation should count for more; rather, this depends on broader philosophical considerations about what kind of a confirmation theory can adequately capture the epistemic principles that (should) govern scientific confirmation.

Second, demanding that confirmation should be purely logical and thereby ruling out predictivism simply begs the question against the predictivist. Even in the case that a purely logical version of Bayesianism turns

[25] Steele and Werndl (2013) make this kind of an argument in the context of climate model confirmation. They argue that based on the Bayesian approach, using evidence both in the construction and in the confirmation of climate models is acceptable.

[26] Glymour (1980) argued famously that Bayesianism is fatally committed to predictivism: after all, the probability of any known piece of evidence is 1. Based on the Bayesian formula, this appears to mean that any known evidence cannot confirm a hypothesis at all. There are many attempted solutions to this 'problem of old evidence.' One approach is to 'excise' the evidence (e) from one's background knowledge prior to applying it in the Bayesian formula, i.e. consider what impact e has on P(H) in the *counterfactual* situation where e is presented as new information (see Howson & Urbach 1996).

out to provide an attractive theory of confirmation, a defense of such a confirmation theory against predictivism should engage with the predictivist position, showing where it goes wrong. We have seen numerous arguments for the view that predictions count for more than accommodations in science. A compelling theory of confirmation that denies the conclusion of all of these arguments should be able to engage with the arguments on more direct terms rather than settle the issue by fiat. (A predictivist could similarly challenge Bayesianism, arguing that it does not take into account the various advantages of prediction over accommodation.)

Another point that can be raised against the principled objection to predictivism concerns the idea that scientific confirmation should be objective. One attractive feature of the purely logical approach is that it appears to deliver in this regard: when contingent considerations such as use-novelty are dispensed with, we are left with an impersonal logic of confirmation. In contrast, predictivism is vulnerable to historico-biographical accidents, where the same piece of evidence can provide different degrees of confirmation depending simply on its contingent role (see Votsis 2014, pp. 75-76). However, here it can be responded that factoring in contingent considerations does not necessarily mean that the resulting theory of confirmation will *not* be objective. Methodological rules and principles in science could very well invoke contingent factors, but this does not mean that these rules cannot themselves be part of an objective scientific process (see Mayo 2014, p. 82). Schurz (2014, p. 91) reminds that in statistics rules about contingent procedures are not unfamiliar at all: the reliability of statistical inference depends crucially on whether or not the data was collected randomly. The inclusion of contingent factors therefore does not automatically mean that we have to abandon the quest for an objective theory of scientific confirmation. On the contrary, if there are contingent factors that are demonstrably associated with more (or less) reliable forms of inference, their inclusion appears to make our confirmation theory *more* rather than less objective (with regard to what is true about the world).

The generalist approach. The generalist approach to anti-predictivism employs a more comprehensive strategy. Here, the goal is, roughly, to expand the picture on the prediction versus accommodation issue. Predictivist arguments are typically pursued with a strategy where the predictivist focuses on a particular advantage with prediction, or a flaw with accommodation, and then proceeds to argue that there is a predictivist advantage in science. However, what is missing is the big picture: how does novel prediction compare to other sources of epistemic support in science? Furthermore, is it really the case that prediction has nothing but advantages over accommodation, but might there also be *dis*advantages?

Harker (2008) provides a general critique of the predictivist position along these lines. Harker rejects strong predictivism, appealing to arguments raised by weak predictivists, and proceeds to then argue that weak predictivist theses themselves are misleading. He argues that weak

predictivists have done very little to show that novel prediction compares favorably to other epistemic indicators in science, such as advocacy of reputable scientists, longevity of support, widespread support among scientists, etc. A novel prediction might in some limited circumstances count for more in confirming a particular theory, but so does simply the support of a reputable individual within the scientific community. Harker continues to make the case that predictivist intuitions can be accounted for by our preference for progress and for theories that explain more with less assumptions. When these preferences are taken into account, Harker (2008, p. 451) holds that "predictivism becomes redundant."

Two points can be raised against Harker on behalf of the predictivists. First, Harker's argument about comparing novel predictive success to indicators such as longevity of support and advocacy of reputable individuals appears subtly to miss the target on the underlying issue in the prediction versus accommodation debate. A chief concern in this debate has been to develop philosophical theories about what underlies rational judgments about theory confirmation in science. That is, the issue is to explain and justify why certain types of evidence should generate the kind of second-order indicators of evidential support that Harker refers to (e.g. longevity of support, widespread support, advocacy by reputable scientists). It may very well be true that practicing scientists place epistemic weight on sociological factors such as the support of reputable individuals and the community at large. However, predictivists could argue that these sociological factors are useful only in so far as they reliably reflect underlying evidential considerations, among which novel predictions are particularly important.[27] The idea that novel predictions are important and fundamental to evidential reasoning in science could be wrong. However, in this case the issue for the anti-predictivist is to address *this* claim and show why novel predictions are not important or fundamental.

The second part of Harker's argument attempts to do just this by appealing to our preference for progress and explanatory strength and simplicity: Harker argues that whatever novel prediction is supposed to achieve is covered by these other evidential preferences. However, I am not sure that Harker has yet done enough to show that predictivism can be dispensed with so easily by referring to these factors. In a similar way as Harker himself argues that predictivists have not provided much evidence about the special value of novel prediction in science, he does very little to show that a preference for progress, strength, and simplicity can account for the myriad of advantages that predictivists have argued for. For example, Harker's (2008, p. 448) argument against Lipton's (2004) fudging theory is simply to restate the claim that prediction is no more reliable indicator of a unified theory than the reputation of the author. In the absence of further

[27] It is, of course, possible to question this entire enterprise and argue that it is sociological factors "all the way down." This is a (much) larger debate beyond this study, and one that Harker (2008) does not appear to be invoking.

evidence from scientific practice, it seems to me that this argument in not compelling in either direction. To be able to evaluate whether the predictivist or the anti-predictivist is more likely to be correct here, we need to be able to engage more closely with the relevant scientific literature on this question (see Chapter 4).

Another recent contribution to what I have called the generalist approach to anti-predictivism is that of Dellsén (forthcoming). In a paper titled "An Epistemic Advantage of Accommodation over Prediction," Dellsén makes the case that what is missing in the debates on prediction and accommodation is an appreciation of *dis*advantages of prediction. Dellsén argues that novel prediction is associated with a so far overlooked problem: in cases of novel prediction, there are more concerns that the data might have been manipulated or fabricated, given that it is new and more uncertain. Accommodated evidence, in contrast, is *less* uncertain, given that it is already known and available in the scientific community. For this reason, Dellsén argues that accommodation is actually (at least in certain cases) better than prediction. A similar point has actually also been raised by Harker (2008, p. 440), and even earlier by Brush (1989, p. 1127), but Dellsén develops the argument in more detail. Dellsén shows that the conclusion that accommodation is superior to prediction can be justified by Jeffrey Conditionalization, where the Bayesian formula is modified to include one's confidence in the evidence itself. Given the uncertainties that surround novelly predicted data, accommodated data that is less uncertain provides stronger confirmation. Dellsén does not directly reject all weak predictivist theories; in fact, he argues that they may also hold in some cases. However, when the respective advantages and disadvantages of both prediction and accommodation are taken into account, Dellsén holds that accommodation may be superior all things considered, or at least there is no general reason to prefer prediction to accommodation in science.

Dellsén's new argument provides, in my view, a much needed new perspective into the prediction versus accommodation debate. A lot of attention has been placed on the virtuous aspects of prediction and the unvirtuous aspects of accommodation, but the possibility of converse advantages and disadvantages has been overlooked. Dellsén provides the first argument that takes this possibility seriously. He does not engage directly with many of the weak predictivist arguments in the literature, so the comparison between the respective advantages and disadvantages of prediction and accommodation is yet incomplete. However, the perspective that Dellsén introduces is worthy of further exploration, a task that awaits us in Chapter 4.[28]

[28] Interestingly, there is also a potential counterargument to Dellsén, introduced earlier in the prediction versus accommodation debate. Akeroyd (2003, pp. 339-340) proposes that novel predictions should be preferred, as due to improvements in scientific methodology, new data is more likely to have been reliably produced than old data. A balanced evaluation of various predictivist arguments is conducted in Chapter 4.

2.8 EVIDENCE FROM THE SCIENCES

Before we conclude the discussion on the prediction versus accommodation issue, there is one more important perspective that needs to be included: that of the scientists. The prediction versus accommodation issue has been investigated in the context of scientific practice by philosophers of science, science historians, as well as the scientists themselves. What can we learn about the issue of predictivism based on the practices and statements of scientists?

This question has inspired a lot of debate within the philosophical community on the prediction versus accommodation issue. One of the earliest case studies that evaluated a philosophical theory about predictivism against detailed evidence from the sciences is Worrall's (1989b) study on the scientific evaluation of Fresnel's wave theory of light in the 19th century. In this famous case of scientific history, the wave theory of light entered the field facing stiff competition from the dominant particle theory of light. However, the wave theory unexpectedly triumphed over its dominant rival after the confirmation of a highly surprising novel prediction. The wave theory implied that if a small opaque disk is held in front of a light emanating from a small hole, the center of the disc's shadow will be bright – a consequence considered absurd at the time (see Worrall 1989b, p. 136). Yet, when this prediction was tested by experiment, it was spectacularly confirmed. This episode appeared, on the surface, to provide a clear illustration of strong predictivism in practice, where the acceptance of a new theory followed the confirmation of a surprising novel prediction (see Giere 1984). However, Worrall examines detailed historical documents on how elite scientists in France reacted to this case, and argues that there was no strong response to the white spot prediction. On the contrary, he finds that the scientists emphasized Fresnel's handling of already known straightedge diffraction cases. Worrall notes that a lengthy evaluation report by leading scientists on Fresnel's theory contains only two sentences on the white spot prediction, and many evaluators continued to resist the wave theory long after the confirmation of this prediction. Worrall argues that this case actually supports his use-novelty account, according to which what mattered epistemically was that facts about the known diffraction cases were not used by Fresnel in the construction of the theory.

Case studies have been employed in the prediction versus accommodation debate regularly ever since Worrall's seminal example. The case that has perhaps inspired the most commentary is that of Mendeleev and his periodic system (see, for example, Maher 1988; Brush 1996, 2007; Scerri & Worrall 2001; Akeroyd 2003; Scerri 2007, pp. 123-157; Barnes 2008, pp. 82-121, 2014; Schindler 2008, 2014). This case may have appeared particularly suitable in that Mendeleev achieved some of the most famous novel predictive successes in the history of science, using his periodic system to derive multiple successful predictions about the existence of new elements and their properties. Here, points of contest have included the timing and

frequency of mentions of Mendeleev's periodic law in scientific journals and textbooks, the comments of contemporary scientists, and Mendeleev's overall predictive record.[29] The science historian Stephen G. Brush (1996), for example, finds that mentions of Mendeleev's periodic law increased sharply in scientific publications in America and Britain after the confirmation of Mendeleev's first novel predictions. Supplementing this evidence with the writings of contemporary scientists, Brush concludes that (temporally) novel predictions did play some role in the acceptance of the periodic system. Scerri and Worrall (2001, pp. 428-436) challenge both Brush's conclusion and evidence, pointing out that even though there was an increase of mentions, well over half of chemistry textbooks published in America, Britain, France, and Germany around the time when Mendeleev's predictions were confirmed fail to mention his periodic law. In fact, even among the textbooks that do mention the law, only just over half mention the novel predictions. They argue that the writings of contemporary scientists fail to show a preference for temporally novel prediction. Another contentious point has been the citation of the prestigious Davy Medal, which was awarded jointly to Mendeleev and Meyer for their work on the periodic system (see Spottiswode 1883, p. 392). The citation makes no mention of Mendeleev's novel predictions, which could be seen as a strike against predictivist ideas (see Scerri & Worrall 2001, p. 416-417).[30]

The largest collection of case studies on the prediction versus accommodation issue is undoubtedly Brush's (e.g. Brush 1989, 1990, 1993, 1994, 1995, 1996). Brush's expansive body of work, developed in the field of science history (but also addressing directly the relevant philosophical debate), reveals an interesting conclusion that any predictivist must contend with. Brush (2007) writes that the Mendeleev case is, in fact, *an exception* where there is evidence that (temporally) novel predictions played *some* epistemic role. In multiple case studies in physics, Brush finds no evidence for predictivism. On the contrary, he finds that scientific theories are sometimes rejected despite their successful novel predictions, and sometimes they are accepted independent of the confirmation of novel predictions. Brush writes that he has not found a single example where novel predictions were crucial to the acceptance of a scientific theory (see Brush 1994, p. 140; Brush 2007, p. 259). From the philosophical point of view, a limitation of Brush's research is that it predates the introduction of more nuanced predictivist arguments into the philosophical literature (namely, the weak versions of predictivism). So, it is unclear what Brush's results imply about these particular theories. However, an overarching conclusion of Brush's research is that (temporally) novel predictions have not been important in many natural scientific fields to scientists themselves (particularly in physics).

[29] For recent discussion about Mendeleev's predictive record, which also included many false predictions, see Scerri (2007), Stewart (2019), Wray (2019), and Lente (2019).

[30] A competing interpretation of this evidence that is more favorable to predictivism is provided by Akeroyd (2003, p. 341) and Barnes (2014, p. 51).

Most recently, the prediction versus accommodation problem has emerged as a contentious issue in the social sciences, sparking a lot of commentary and debate (see, for example, Kerr 1998; Bosco et al. 2016; Shaw 2017; Hollenbeck & Wright 2017; Rubin 2017, forthcoming; Vancouver 2018; Murphy & Aguinis 2019). Kerr (1998) introduced into the social scientific literature the issue of 'HARKing,' i.e. 'hypothesizing after results are known.' HARKing means presenting post hoc hypotheses as if they were a priori hypotheses; i.e. in our terms, presenting accommodated hypotheses as use-novel hypotheses. According to Rubin (2017, forthcoming), interest in Kerr's paper begun to increase sharply following what is called the (still ongoing) 'replication crisis,' where multiple prominent research findings in the social and medical sciences have failed to replicate in subsequent studies. Scientists have engaged in an active debate about HARKing and its possible role in contributing to low replication rates in their fields. Commentaries range from what can be interpreted as presenting predictivist views (e.g. Bosco et al. 2016; Shaw 2017) to those that are more skeptical about the alleged dangers of HARKing (e.g. Vancouver 2018; Rubin forthcoming).

All of these debates remain open without generally accepted conclusion. In the philosophy of science, important cases such as the Mendeleev case are still being contested without a solution to whether predictions (and what kind of predictions) counted for more than accommodations. The most comprehensive collection of case studies in the history of science (by Brush) provides evidence against stronger forms of predictivism, but it is yet limited to the natural sciences, and does not directly address the myriad of weak predictivist arguments that have appeared in the philosophy of science. In scientific practice, it is not difficult to find both predictivist and anti-predictivist sentiments expressed by individual scientists (see, for example, Brush 1996; Scerri & Worrall 2001).[31] The most recent scientific debate on prediction and accommodation in the social sciences, rather than providing a unified view of scientific opinions about the confirmatory value of novel predictions, shows that the issue is unresolved even among currently practicing scientists. In fact, the scientists have even turned to the contemporary literature on the prediction versus accommodation issue in the philosophy of science in looking for a solution (see, in particular, Rubin 2017).

Whether we refer to philosophical case studies, historical case studies, or contemporary debates in science, it appears fair to say that the

[31] The physicist Ne'man, for example, writes that "…the importance attached to a successful prediction is associated with human psychology rather than with scientific methodology. It would not have detracted at all from the effectiveness of the eightfold way if the Ω^- had been discovered before the theory was proposed. But human nature stands in great awe when a prophecy comes true, and regards the realizations of a theoretical prediction as irrefutable proof of the validity of the theory." (Ne'eman & Kirsch 1986, p. 202) The chemist Wilson (1952), in contrast, writes that "[s]uccessful prediction is usually considered stronger support for a hypothesis than the explanation of an equal quantity of observation known to the creator of the hypothesis at the time of its creation. This is not hard to justify on purely logical grounds and appears valid as a result of experience[.]"

evidence from scientific practice has not *settled* the prediction versus accommodation issue. Given the numerous arguments (and contradictory intuitions) both for and against predictivism, as well as the various different versions of predictivism (strong, weak, temporal, use-novel, etc.), this is perhaps unsurprising. Case studies and evidence from the sciences have an important role to play in clarifying this issue, but they can only go so far with such a heterogeneous field of alternatives. A satisfactory resolution of the prediction versus accommodation issue will require a more balanced investigation, which takes into account both predictivist and anti-predictivist views, their sometimes conflicting starting points and assumptions, and the relevant evidence that has arisen in scientific practice. The present study attempts to contribute to such an investigation, engaging with both the philosophical and the scientific literature on this issue.

2.9 SUMMARY

We have now explored contemporary arguments and views on the prediction versus accommodation issue. It is time to draw overall conclusions about where the discussion stands:

1) Strong versions of predictivism, where novel predictions are considered to have an intrinsic, global advantage over accommodation, are no longer considered compelling in the prediction versus accommodation debate. Both predictivists and anti-predictivists agree that the issue of prediction versus accommodation is more nuanced, where the advantage of one over the other depends more closely on the circumstances. Strong predictivism also appears questionable in light of evidence from scientific practice.

2) Contemporary predictivists in the philosophy of science defend a variety of weak predictivist views. The methodological approach to weak predictivism employs a *negative* strategy in supporting the special value of novel prediction. It holds that accommodation is associated with unreliable methodologies and methodological choices (e.g. overfitting, fudging, and hypothesis hunting). The advantage of novel prediction is that it allays concerns about the possible use of these unreliable methods. The agent-based approach employs a *positive* strategy on the behalf of predictivism. It holds that novel prediction provides an epistemic boost for the theory because it speaks to the reliability of the theorist. Finally, the logical approach argues that prediction and accommodation should be understood in a logical rather than contingent sense. Logical prediction provides full support to the theory, while logical accommodation has less epistemic force.

3) Many of the weak predictivist theories are compatible, and they could be seen as providing complementary answers to why novel predictions confirm the theory more strongly than accommodations. It is possible to endorse both the methodological and the agent-based approach to predictivism (see Barnes 2008). A pure version of the logical approach rejects

the inclusion of contingent factors in confirmation theory, so it is incompatible with the methodological and the agent-based theories. However, the logical approach can also be expanded with methodological rules (see Schurz 2014).

4) Anti-predictivists argue that the advantages of novel prediction that predictivists postulate can be accounted for (largely or completely) by other epistemic factors or principles. There is nothing particularly special about novel predictions. Furthermore, novel predictions have certain *dis*advantages that may make accommodations superior overall.

5) The prediction versus accommodation issue is contested and remains unresolved even among contemporary scientists. Scientists as well as philosophers have expressed a variety of both predictivist and anti-predictivist views.

Before we move on to evaluate predictivism in its various forms, we must consider one more facet of the debate: the issue of scientific realism. The debate on scientific realism has proceeded independently from the prediction versus accommodation debate, but novel prediction plays a very prominent role in this debate, making the two debates intimately connected. Predictivist views have often been developed with the ultimate goal of defending scientific realism against the anti-realist challenge. In the context of the realism debate, *the positive, explanatory strategy* for defending predictivism and realism is studied more closely.

3 SCIENTIFIC REALISM AND THE NO MIRACLES ARGUMENT

One of the most important purposes of the analysis of novel prediction is to apply it in the defense of scientific realism. The debate on realism and anti-realism is among the most central in the philosophy of science. On it turns the very existence and nature of scientific knowledge. The scientific realist holds a positive epistemic attitude about scientific theories and models: scientific theories and models are successful in describing both observable and unobservable aspects of the world, and we are justified in believing in the truth of their contents. The scientific anti-realist typically agrees with the realist about the observable aspects of science. However, she questions whether belief in *un*observables is warranted. The anti-realist argues that we should believe only that scientific theories are 'empirically adequate'; i.e. theories successfully describe the observable world, but beyond that we should remain skeptical or agnostic.

In the past few decades, the scientific realism vs. anti-realism debate has been largely a tale of two arguments. On the side of scientific realism, there is the 'No Miracles argument' (NMA), which holds that the success of science would be "a miracle" unless science is successful in approximating the truth (see Putnam 1975, p. 73). In other words: unless we attribute truth to scientific theories, how on earth do we explain their immense success? The kind of success that counts in the NMA is now commonly specified to be novel predictive success, as the accommodation of empirical results does appear to be explainable without attributing truth to the theory (see previous chapter). The realist believes that the best, if not the only, explanation for novel success is that the theory has latched on to the truth (e.g. Psillos 1999, pp. xx-xxi). On the side of anti-realism, there is the 'Pessimistic Meta-Induction' (PMI) argument. Introduced by Laudan (1981), this argument points to the history of science, where many theories that were once considered empirically successful have now been abandoned. Given the long history of refutations, it appears possible, and perhaps even likely, that any of the theories that realists currently believe worthy of realist commitment will ultimately suffer the same fate as their predecessors, and hence we should refrain from committing to their truth.

The overall debate between scientific realists and anti-realists spans beyond these two arguments. However, here we will focus on these arguments, as they are particularly relevant in considering the role of novel predictions in supporting scientific realism. They are also rather important within the overall dialectic; for example, Chakravartty (2017a) describes the No Miracles argument as "[t]he most powerful intuition motivating realism[.]" Other aspects of the scientific realism vs. anti-realism debate are touched on briefly to help explain the role of the NMA and the PMI within the overall

debate. The goal of this chapter is to develop an understanding of the novel prediction based defense of scientific realism and its role in supporting the realist view of science.

The chapter proceeds as follows. In section 3.1, I describe the contemporary realist position, and briefly go over different varieties and aspects of the realist view of science. I contrast the realist position with contemporary scientific anti-realism, in particular the constructive empiricist variant that has been most influential in the philosophy of science. In section 3.2, I discuss the novelty-based NMA. Here, we see more closely how the realists have developed the positive, explanatory strategy to support the special value of novel prediction (cf. section 2.5). Important specifications to this strategy are discussed, and at the end of the section I compare the NMA to other recent arguments for scientific realism. Section 3.3 explores the PMI, while other central anti-realist arguments are also briefly introduced. Finally, in section 3.4, I provide an overall conclusion about the novelty-based defense of scientific realism and the prediction vs. accommodation issue. This serves as the basis for subsequent chapters of the study, where the epistemic value of novel prediction in science is evaluated.

3.1 SCIENTIFIC REALISM AND ANTI-REALISM

Scientific realism is a rich and multi-faceted philosophical position that concerns the nature of the scientific enterprise. There is no single thesis or creed that characterizes the entire position. Chakravartty (2017a) remarks that "[it] is perhaps only a slight exaggeration to say that scientific realism is characterized differently by every author who discusses it[.]" However, there are certain common core ideas. Psillos (2000, p. 706) offers the following theses as constitutive of scientific realism:[32]

The metaphysical thesis. The world exist independent of our ability to describe or observe it. It has a definite, mind-independent structure.

The semantic thesis. Scientific theories provide truth-conditioned descriptions of the world. That is, theoretical terms have putative factual reference (they are capable of being true or false).

The epistemic thesis. A belief in the approximate truth of appropriately successful (i.e. 'mature,' 'novelly successful') scientific theories is warranted.[33] Despite the fact that the world has a mind-independent

[32] Another common idea concerns the *aim* of science: a realist can be characterized as someone who believes that the aim of science is to provide true descriptions of the world (see, for example, van Fraassen 1980). This is undoubtedly one belief that characterizes a realist attitude towards science, but for many contemporary realists it has appeared too modest to constitute a worthwhile realist stance on its own. Such a view, after all, is compatible with science never achieving truth at all. Most realists hold a more substantive commitment in terms of *the achievements* of science (see Chakravartty 2017a; Rowbottom 2019, pp. 453-459).

[33] Realists sometimes use the somewhat vague term 'mature' to describe the appropriate kind of successful scientific theories (see Chakravartty 2017a). According to Chakravartty (2017a), maturity can

structure, we are capable of reaching epistemically justified beliefs about what that structure is.

The semantic thesis was the subject of much controversy in the 20th century, particularly in relation to the logical empiricist program in philosophy. According to Psillos (2000, p. 707; see also Chakravartty 2017a), semantic realism is no longer contested. Scientific theories are considered to contain indispensable theoretical terms that refer to things beyond what is immediately observable. Naturally, this is not to say that all theoretical terms are always to be interpreted literally as either true or false (see Rowbottom 2019b, pp. 259-261). Scientific theories also contain abstractions and idealizations that deviate from what is literally true about the world. However, most contemporary realists and anti-realists have agreed that scientific theories *can* make claims about unobservable theoretical mechanisms, laws, or entities that are capable of being true or false.[34]

The metaphysical thesis stands in contrast to traditional idealist or phenomenalist views of reality and to modern verificationist accounts where no distinction is drawn between the world and what is believed to exist in it based on appropriate epistemic practices. In the past few decades, the anti-realists within the philosophy of science have generally agreed with the realist on the metaphysical thesis (along with van Fraassen 1980). The issue for the realist has been to defend the *conjunction* of the metaphysical and the epistemic theses (see Psillos 2000). If reality genuinely has mind-independent structure, how is it possible for us to reach justified beliefs and knowledge about that structure? The scientific anti-realist questions whether we are (or can be) in a position to achieve such warrant for scientific theories. The realist argues that appropriate epistemic justifications can give us this warrant (e.g. the No Miracles argument).

Some further distinctions within the realist position can be drawn based on what, exactly, realists are realists about. There are a number of more modest variants of realism where what is taken to be real is restricted significantly. *Entity realism*, for example, claims that we should believe in the existence of unobservable entities that scientists manipulate, but not in the theories that describe those entities (see Hacking 1983). *Structural realism*, inspired by Poincaré and rehabilitated in modern times by Worrall (1989a), argues (roughly) that we should believe in the mathematically describable relations between entities but not in their 'natures.' Epistemic versions of structural realism hold that even if entities have a real nature, justified belief is restricted to structural relations, while stronger, ontic versions of structural realism argue that structures are either more basic than entities or all that

be understood in terms of the ability of the scientific theory to make novel predictions. Psillos (1999, p. 102) writes that maturity "can be characterised by the presence of a body of well-entrenched background beliefs about the domain of inquiry which, in effect, delineate the boundaries of that domain, inform theoretical research and constrain the proposal of theories and hypotheses."

[34] An exception to this may be Rowbottom (2019a), who defends a new instrumentalist view of science.

there is to begin with (see Frigg & Votsis 2011 for a review).[35] With the wide range of options available, it can also be asked whether any form of realism should be applied to the whole of science, or rather different areas or fields of science viewed in different ways (see Asay 2019). For example, one could be a realist about biology but anti-realist about fundamental physics.

A few words are in order with regard to the novelty-based defense of scientific realism and the different, more restricted versions of realism. The novelty-based defense of realism has typically been associated with a more substantive form of realism that recommends belief in the actual contents of scientific theories, i.e. the theoretical laws, mechanisms, and entities described by scientific theories (e.g. Psillos 1999), or at least sees no reason to exclude them from the realist position (see Vickers 2019, pp. 577-578). Indeed, this view can be defended against the more restricted versions of realism by arguing that factors such as structure alone cannot explain the novel predictive success of scientific theories (see Psillos 1999, pp. 147-148). Structural realists, in turn, have generally not appealed to novel predictive success as a criterion for identifying the structural relations that we should have confidence in (see Frigg & Votsis 2011; although see also Worrall 1989a, 1989b). On the other hand, the novelty-based defense of realism is by no means incompatible with more restricted forms of realism such as the structural realist approach (see Vickers 2019). One could appeal to novel predictive success as an important epistemic indicator in science, but hold that what it speaks to are real structural relations rather than the real 'natures' of entities in the world. For this reason, it is useful to identify what Vickers (ibid.) has recently called 'neutral realism.' The neutral realist does not take a stance on the question of whether scientific knowledge is ultimately structural or not, but recognizes this as one possibility. In other words, the neutral realist may believe in more than structural content, but is open to the possibility that a structural understanding of scientific knowledge could ultimately prove most satisfying.

In what follows, I take it that one can be a scientific realist in the neutral, inclusive sense. A scientific realist is someone who holds that we can obtain justified beliefs about the reality of the unobservable theoretical mechanisms, laws, and entities that feature in scientific theories (for the purposes of prediction and explanation) (e.g. Psillos 1999; Vickers 2019). Whether theoretical mechanisms and laws should ultimately be understood in structural terms (or entity-talk replaced by talk about structures) is one possibility that continues to be explored in the philosophical literature on alternative forms of realism (cf. Vickers 2019).

In contrast to the scientific realist, the scientific anti-realist is someone who recommends that we should be more skeptical about the contents of scientific theories and models. The contemporary anti-realist program in the philosophy of science is deeply associated with Bas van

[35] Given the significant concessions these positions make to the anti-realists, they could also be described as alternatives to realism rather than versions of it (see Rowbottom 2019b).

Fraassen's (1980) *constructive empiricism*. The constructive empiricist grants to the realist that the world has a mind-independent structure, and even agrees that we can attain justified beliefs about the observable world based on our senses. However, she questions whether we are justified in believing in what scientific theories say about the *un*observable aspects of the world. In other words, the constructive empiricist questions (most of) the theoretical knowledge that science is supposed to provide. As an alternative to the realist's truth-based explanation for the empirical success of science, the constructive empiricist offers *empirical adequacy*. We should believe that scientific theories succeed in 'saving the phenomena:' they correctly describe observable phenomena and relations in the world, but what they say about unobservables beyond that may very well fail to represent the real nature or structure of reality.

The crucial feature that characterizes the constructive empiricist variant of scientific anti-realism is epistemic agnosticism (or skepticism). The scientific anti-realist holds that there is no compelling reason to extend our commitment to scientific theories beyond a belief in their empirical adequacy. The scientific anti-realist does not argue that the contents of scientific theories are definitely false. Rather, the issue concerns epistemic justification: we lack the warrant for belief when it comes to unobservables in science. In one of the most recent defenses of scientific anti-realism, Wray (2018, p. 71) emphasizes this feature of anti-realism, writing that he thinks that "even the most skeptical contemporary anti-realists would think that it is doubtful that all the theoretical entities postulated more than fifty years ago do not exist." The gist of the anti-realist position is to argue simply that realists are overreaching with their epistemic commitments: scientific theories are not on a secure enough basis for us to have warranted belief in their truth.

Even though van Fraassen's constructive empiricism is very much the prototype, contemporary scientific anti-realists too vary in what they are anti-realists about. One way to modify the anti-realist position concerns the distinction between observable and unobservable aspects of nature. This distinction may not always be clear, and an anti-realist could also draw the relevant distinction between what we should and should not believe in a different way. Stanford (2006, pp. 27-37), for example, holds that although observability is important, the more fundamental divide runs between cases where we can and cannot reliably exhaust relevant alternatives to our theories. He argues that we should remain anti-realist about more fundamental scientific theories and hypotheses that reach farther into the unfamiliar (where we have less basis to rule out alternative explanations), and not unobservables per se.[36]

For the purposes of what follows, I take it that scientific anti-realism covers positions such as that of van Fraassen (1980), Stanford (2006),

[36] Stanford is a realist, for example, about dinosaurs and microscopic entities, but anti-realist about more fundamental theories and hypotheses such as common ancestry, the nature of chemical bonds, Einstein's theory of gravity, etc. (see Stanford 2006, pp. 33-34)

and Wray (2018). The anti-realist is someone who questions the realist's commitment to theoretical knowledge in science; e.g. the unobservable entities, mechanisms, and laws that scientific theories postulate. We can and should believe that scientific theories are empirically adequate, but we should refrain from extending our commitment to their truth. It is possible for one to be an anti-realist and still allow for *some* theoretical knowledge in areas where the evidence appears particularly exhaustive of relevant alternatives (e.g. Stanford 2006). However, for the most part, the anti-realist holds that we should remain agnostic or skeptical about the unobservable contents of scientific theories.

3.2 THE NO MIRACLES ARGUMENT

Let us then move to discuss the novel prediction based defense of scientific realism, i.e. the novelty-based No Miracles argument (NMA). The original statement of the NMA is typically attributed to Putnam (1975, p. 73), who argued that realism is the only philosophy that does not make the success of science a miracle (cf. Whewell 1840). In other words, realism is the only philosophy of science that provides a *good explanation* for the success of science. Anti-realism, in contrast, appears to make the success of science into some type of cosmic coincidence. The clarification that the success of science should be understood in terms of its *novel* success arose particularly in response to Laudan's (1981) original article on the Pessimistic Meta-Induction argument, where he showed that many theories that were once considered empirically successful have now been abandoned. The realists responded that Laudan is using too broad a notion of empirical success. There is nothing particularly surprising when theories that were developed about known phenomena are abandoned, as they were designed to account for those phenomena in the first place (e.g. Musgrave 1988). What is remarkable, and indeed in need of explanation, is when theories successfully predict novel facts that were not used in their construction – only these should count for the NMA. In the years after, the realists have further specified their argument, and there has also been discussion about how the argument should be formulated more precisely. I will start with the important specifications, and then move on to discuss the form of the NMA as a philosophical argument, its criticism, and its role in the defense of scientific realism.

Specification 1: 'surprisingness.' The first specification to the novelty-based NMA concerns the type of novel predictions that theories achieve. It is now generally recognized that predictions, even if (use-)novel, are not always very impressive. A prediction of an empirical result can be strictly novel, in that the result was not used in the construction of the theory, but it may concern something that was already expected or easy to get correct by accident. Vickers (2013, pp. 195-196) gives the example of Immanuel Velikovsky's catastrophist theory of the Solar System. According to this theory,

the Earth was subject to catastrophic close contacts with other planets in the Solar System in historical times. The theory is fantastical, but Velikovsky used it to derive a successful novel prediction: Venus is "hot" rather than "cold." Vickers argues that the realist should not be impressed, because the prediction is so vague. We are, in effect, talking about a 50/50 chance of success irrespective of the truth of the theory. If a novel prediction is to count for something in the NMA, it should do more than simply make a vague statement that could easily turn out correct by accident.[37]

To rule out cases like Velikovsky's, the novelty-condition is specified with terms such as 'bold' (Barnes 2008, p. 186), 'impressive' (Vickers 2013), 'surprising' (Carman & Diez 2015), 'risky' (Vickers 2019), etc. Unfortunately, not much effort has been spent by realists in making this qualifier more precise. A qualitative description has been considered enough to distinguish between cases that clearly count as impressive novel successes (e.g. Einstein's general relativity and the bending of light) and cases that do not (e.g. the Velikovsky case) (e.g. Vickers 2013, p. 195). As one possible specification, Vickers (2019, p. 580, fn. 22), referring to Hájek and Joyce (2008), suggests that some realists could use a Bayesian measure of "surprisingness," where the less probable the empirical result appears from the point of view of an evaluator, the more it confirms the theory. In other words, a quantitative measure of the surprisingness of novel success could be introduced, where the confirmatory force of the novel success depends on how surprising it is to the evaluator.

With the surprisingness or impressiveness qualifier, the realist's position has become somewhat more complicated. Stanford (2009, p. 384) argues that the new specification raises a threshold problem, where the realist is unable to provide a principled criterion for how to identify the cases of novel success that *are* surprising or impressive enough to warrant realist commitment. Vickers (2013, p. 197) responds that providing a threshold is not an issue as far as the realist position is understood to incorporate degrees of confidence. Realist commitment can comes in degrees, and so the more surprising or impressive the novel success the greater the realist commitment. In what follows, I will recognize the surprisingness qualifier as an underlying condition for the NMA, so that 'novel success' is understood to mean impressive or surprising novel success (along the lines specified by Vickers 2013 and Vickers 2019). That is to say, as far as we are evaluating the viability of the novelty-based criterion of realist commitment, we need to pay attention to *surprising* or *impressive* novel success rather than just novel success per se.

Specification 2: the 'divide et impera' strategy. Another important specification that realists now make is to go 'selective,' i.e. they

[37] Vickers (2013, p. 196) emphasizes that the problem is not that Velikovsky's prediction is qualitative rather than quantitative. He points out that qualitative predictions can be just as impressive if they pertain to a novel phenomenon that could not feasibly have been conceived without the help of the theory.

restrict their realist commitment only to certain *parts* of scientific theories and models. We have already highlighted a few selective strategies (e.g. entity realism and structural realism). However, in the context of the novelty-based argument for scientific realism, this move is associated with a particular kind of selectivity, introduced by Kitcher (1993, pp. 140-149) and Psillos (1999, pp. 103-109). Kitcher and Psillos observe that in cases where scientific theories make successful novel predictions, not all constituents of the theory necessarily play a role in the derivation of those predictions.[38] In Kitcher's terms, some theoretical parts or constituents may be merely 'presuppositional' (or 'idle'), while some are 'working' and genuinely contribute to the theory's predictive successes. In cases of novel predictive success, Kitcher and Psillos thus argue that the question should *not* be about the theory as a whole, but rather its *constituent parts*. Realist commitment must be reserved only to those constituents that really 'fuel the derivation' of the successful novel predictions.

This specification has been called the 'divide et impera' strategy (see Psillos 1999), or alternatively 'deployment realism' (see Lyons 2006). It is specifically motivated by the historical challenge raised by Laudan (1981). Along with the move to emphasize surprising novel success, the introduction of this strategy is one of the most important defenses that realists have put forward to support the NMA against the PMI. With the more nuanced understanding of scientific theories and their constituent parts, any abandoned theory that at some point achieved novel predictive success does not automatically create a counterexample to realism. Instead, what counts is *continuity*. Even if a novelly successful theory is abandoned in the course of scientific practice, certain parts or components of the theory may carry over and be incorporated into later theories. If those parts or components were the constituents responsible for the original novel predictive successes of the abandoned theory, the fact that there is theory change poses no problem for scientific realism: scientists have simply abandoned the 'idle wheels' of the older theory and retained the parts that actually latched on to reality (as revealed in their role in the derivation of the novel predictive successes of the theory).

The divide et impera strategy has intuitive force, and it has been considered a prominent tool in the realist's arsenal even by philosophers of science more skeptical about the reach of selective realism (see, for example, Lyons 2017). It is not difficult to find intuitive examples of abandoned theoretical posits that were not needed for the predictive successes of past abandoned theories. For example, Newton's theory included postulates such that the center of mass of the Universe is at absolute rest, which need not be invoked to derive predictions based on Newton's laws (see Psillos 1999, p. 104). If the predictive successes of abandoned scientific theories really did not

[38] An important difference between Kitcher's (1993) and Psillos's (1999) proposals is that Kitcher is concerned with the reference of theoretical terms, while Psillos considers which theoretical constituents scientists themselves consider working rather than idle (see Psillos 1999, pp. 106-107).

depend on false theoretical constituents, the realist has in hand a powerful refinement of the No Miracles argument: the best explanation for the (surprising) novel success of a scientific theory is that the *theoretical constituents* that were deployed in the derivation of that novel success are true.

Even if the intuition behind the divide et impera strategy is clear, things get more complicated when it comes to the details. The realist now needs to be able to provide criteria based on which actual working posits can be distinguished from mere idle wheels, or the realist position threatens to become trivial (e.g. "we should believe some unspecified parts of past and current theories"). Let us consider a few possibilities for such criteria. One alternative that quickly suggests itself is to appeal to scientists' judgments about what is working or idle (see Psillos 1999). However, anti-realists have argued in response that scientists have very often made incorrect evaluations, which simply takes us back to the PMI: contemporary scientists could be wrong similarly as past scientists (see Stanford 2006; Vickers 2013, p. 207). Another possibility might be to identify working and idle constituents from the point of view of a new theory that has superseded an old one. But, as pointed out by Stanford (ibid. pp. 167-80; 2009, 385-87; cf. Vickers 2013), this too amounts to giving up what is important to the realists: we will always be waiting for the next theory to reveal what is actually true about our current theories.

Psillos (1999, p. 105) suggests a criterion of 'availability' to identify genuinely contributing working posits: a theoretical constituent C is indispensable to the derivation of a prediction P if other constituents C' (along with auxiliary assumptions A) cannot yield P and C is not replaceable without loss in the derivation of P by another available constituent C* that is consistent with C' (and A). Psillos's criterion is perhaps suggestive, but it also does not appear to be enough for the realist's purposes. For one, it is not clear what 'availability' will amount to here (see Vickers 2013, p. 202), and as argued convincingly by Lyons (2006, pp. 539-541), Psillos's account appears to defeat the purpose of the selective realist strategy to begin with: the question of whether a particular constituent C contributed to a particular predictive success is independent of the question of whether there are other ways to derive that success. What is needed is a more direct way to give credit where credit is due, i.e. pick out the constituents of the theory that were really deployed in deriving its predictive successes.

A more recent and sophisticated account is provided by Vickers (2013). Vickers draws a distinction between 'derivation internal posits' (DIPs) and 'derivation external posits' (DEPs). DEPs are posits that merely influence or inspire scientists in generating a theory that entails a particular novel prediction, while DIPs are posits that are genuinely deductively connected to the prediction (via appropriate auxiliaries). Vickers suggests that the DEPs, unconnected as they are from the predictions, do not deserve realist commitment. He then argues further that even some of the DIPs can be eliminated (or weakened), as they may 'contain' another posit that is the actual

working part. 'Contain' here means simply that there is some weaker proposition that can be inferred from the original DIP without loss in the derivation of the prediction. For example, the weaker proposition 'the passengers are too heavy' is logically contained within the stronger proposition 'the passengers are 50 kg too heavy' (see Saatsi 2005a, p. 532). Vickers holds that DIPs should be weakened to their minimal logical consequence that has the same deductive consequences with regard to the particular prediction in question. In order to identify actual working posits of the scientific theory, we then work backwards from a particular novel prediction, identify the working parts of the DIPs that were necessary to derive that prediction, and retain those parts in the deductive chain 'upstream' from the prediction that are needed to derive the prediction.

Vickers's (2013) proposal appears to me the most promising one available at this moment. Vickers's account contains the resources to discount a good number of idle posits from scientific theories, which leaves less room for the anti-realist's PMI. The proposal does face challenges as well, which Vickers himself recognizes on multiple occasions. For one, the distinction between DIPs and DEPs may not always be clear: whether we count a certain posit as one or the other may depend on further 'implicit posits' that are needed to derive deductive consequences from particular posits (see Vickers 2013, p. 201). Vickers also concedes that his proposal may still fall short from identifying actual working posits. However, he argues that it can at least be used in identifying idle posits, and realist commitment is then reserved to whatever remains of the theory after the appropriate eliminations have been made. Vickers (2017) continues to press the point against the anti-realist that this will be enough for the realist's purposes. He argues that the onus is on the anti-realist to identify false posits that supposedly featured in novel successes, and it suffices for the realist if she can show in a principled way why those posits were not worthy of realist commitment.[39]

The debate on how to refine the selective realist strategy remains open in the philosophy of science. An important goal for the realist is to produce selection criteria that avoid both *overinclusion* and *underinclusion* of theoretical posits (see Peters 2014, p. 385). In the former case, if it turns out that all kinds of false posits really featured in an uneliminable role in past novel predictive successes, the divide et impera strategy is shown unsuccessful, and the anti-realist wins (see Psillos 1999, p. 105). In the latter case, if the criteria for inclusion are too strict, much of the novelly successful theories may be eliminated, and there may be not much left over what the anti-realist already accepts (e.g. the empirical results as such). Another important

[39] Other selective realist accounts have been formulated by Harker (2013) and Peters (2014). Harker argues that we should reserve realist commitment to those parts of scientific theories that generate empirical progress over previous theories. Peters introduces a new 'empirically successful subtheory account' (ESSA), according to which theoretical posits that unify lower level posits on the way to the derivation of a successful prediction are the posits that should be retained by the realist. Both Harker and Peters deny that (contingent) novel predictions should be invoked in the selective realist strategy.

issue is the question of *prospective* versus *retrospective* identification of working and idle posits, i.e. identifying the relevant posits in current versus abandoned theories. Stanford (2006, pp. 173-180) has argued that if all that the realist can achieve are retrospective identifications, realism is bankrupt, because the whole point of the realist position is to show which aspects of current theories we should believe in. Vickers (2017) responds that prospective identification is at least sometimes possible with regard to idle posits. It is also worth recognizing that *principled* retrospective identification is by no means an easy task, and could be considered an impressive achievement for the realist even if prospective identification were to remain elusive (cf. Peters 2014, p. 380).

The form of the argument. We have now introduced two specifications to the NMA that are commonly endorsed by contemporary realists: surprisingness and selectivity. Next, we should consider the nature of the NMA as a philosophical argument. What kind of an argument is it, and how does it work in supporting realism?

In the course of this study, we have actually already encountered two alternative versions of the NMA. One of these connects novel success to the truth of the theory directly, i.e. *strong predictivism*, and another appeals to the reliability of the theorist in generating the novelly successful theory, i.e. *agent-based weak predictivism*. Strong predictivism, as noted before, has fallen out of favor in the prediction versus accommodation debate. Weak predictivists argue that a direct inference from novelty to truth is fallacious, as facts about the theorist's use of evidence remain unconnected to the inference about the truth of the theory (e.g. Barnes 2002a; White 2003). However, in the context of the scientific realism debate, it is still possible to find statements of the No Miracles argument that apparently take the strong predictivist route. A recent one is Vickers (2019), who defends what he calls the realist's 'success-to-truth' inference. Vickers argues that the realist should commit to the working posits of theories that achieve surprising novel success. In other words, he argues directly that we should infer theoretical truth from novel predictive success in science.

If the weak predictivists are correct, Vickers's argument is fallacious, and it should be abandoned. However, on the behalf of Vickers, it could be suggested that there may be a way in which his argument can be rescued. I propose that this could be done by connecting the argument to *the methodological approach* to weak predictivism. Something like this has already been suggested by Lipton (2004, pp. 181-182), who argues that his fudging account on the advantage of novel prediction over accommodation can explain why accommodation prevents the truth-explanation for the empirical success of the theory.[40] The truth-explanation is available in the case of novel success, but not in the case of accommodation, because accommodation is associated with unreliable methodological practices. Vickers could perhaps

[40] Lipton (2004) himself expresses reservations about the truth-based explanation overall.

appeal to this strategy in maintaining that only in the case of successful novel predictions is an inference to the truth of the theory warranted: accommodation is unreliable, so the success-to-truth inference must be reserved only for novel predictive success.

The second alternative is, of course, the agent-based approach. However, this option may be less attractive in that the agent-based approach was found to suffer from multiple issues earlier (see section 2.5). There are also further options that have been developed within the realism debate. Psillos (1999; 2006; 2011) argues that the NMA should proceed via *a defense of the reliability of the abductive methodology of science*, i.e. the derivation of theories and hypotheses via inference to the best explanation (IBE) in science (cf. Lipton 2004). Psillos's argument has two parts. The first addresses the explanation of the novel success of scientific theories as such. Psillos argues that (Ai) the reliability of first-order scientific methodology, i.e. its ability to yield successful novel predictions and explanations, stands in need of explanation. He argues that (Aii) the best explanation is that the theoretical statements that are implicated in scientific methodology to yield particular predictions are true (see, in particular, Psillos 2011). The second part concerns the reliability of scientific inference. (Bi) The true, predictively successful theoretical statements have typically been arrived at by means of IBE. Hence, we can infer (Bii) IBE is reliable. In other words, Psillos moves to defend the reliability of the NMA (as an IBE-argument) by appealing to the success that has been achieved in science through reliance on IBE.

Psillos's argument is obviously (rule-)circular, which he freely acknowledges: the reliability of IBE is defended by appealing to its reliability in producing results. However, Psillos argues that this argument can be maintained in a broadly externalist, naturalist framework of epistemic justification, where what counts is whether or not IBE really *is* reliable.[41] Psillos explains that the point is to explain and defend IBE to those who employ it, and perhaps sway somebody with a neutral attitude to the realist side, rather than persuade a committed opponent of scientific realism (see Psillos 2011). In this way, the NMA is meant to be shown justifiable within the scientific realist framework.

Both Vickers's and Psillos's versions of the NMA appear to leave in some sense more to be desired. Vickers postulates directly that truth is the best explanation for novel success, while Psillos makes the conditional claim that the NMA is justifiable if IBE really is reliable (in an externalist, mind-independent sense). However, here it should be recalled that the NMA is also backed in an important way by *intuition* (see Chakravartty 2017a; Chapter 1). It seems, intuitively, that only true theoretical constituents can explain the impressive and surprising novel success that we see in science. Worrall (2011) argues that given that we ultimately come back to intuitions (which is not at

[41] Psillos (1999, p. 77) emphasizes that the claim about the reliability of IBE is an empirical one, and if true, only contingently so.

all rare in philosophy), it may be asking a bit too much to formulate the NMA as an argument to begin with. The NMA, for him, is simply an intuition that we "feel" about particular scientific theories: upon observing impressive novel successes, we feel that the theory must somehow have latched on to the "deep structure of the universe" (ibid. p. 21).[42]

Let us take stock. We may summarize the NMA, as presented by scientific realists, as follows. 1) The NMA, whether it is an argument or simply an intuition we feel about scientific theories, is *abductive*: it seeks to provide *an explanation* for the impressive novel successes we see in science (see also Psillos 2006). The realist argues that the impressive novel successes of science require some explanation, and the realist explanation is the best one. 2) The NMA is meant to justify belief in certain outputs of the scientific process, i.e. ultimately the unobservable theoretical laws, mechanisms, or entities that feature in scientific theories to yield particular novel predictions. Even in the versions of the NMA where we take a detour by referring to the reliability of the scientific process or to the reliability of the theorist, the ultimate purpose of this argument is to enable the realist to point to particular scientific outputs and identify them as worthy of realist commitment. In what follows, I will understand the NMA in this broad way, where the realist could proceed with Vickers (2019), Psillos (1999), White (2003), or Barnes (2008), or simply appeal to the intuition that truth (of working theoretical constituents) is the best explanation for surprising novel success in science (Worrall 2011).

Challenges to the argument. The NMA, as a central argument in the philosophy of science, has been the subject of its fair share of critique. A well-known objection, raised originally by Howson (2000), holds that the NMA commits *the base rate fallacy* (see also Magnus & Callender 2004; Howson 2013). The base rate fallacy is familiar in the context of Bayesian reasoning, where we update our beliefs based on our prior beliefs and new evidence that comes in. A standard example is inferring the probability that a particular person in a population has a certain disease. Suppose that we have a test for a disease, Z, that gives a positive result 99 % of the time when the person actually has the disease, and never gives a false negative (i.e. if the person tests negative, the probability that they have the disease is 0). In such a setting, people tend to infer that if a particular person tests positive for Z, the probability that the person has the disease is very high (i.e. 99 %). However, to make such an inference is to commit the base rate fallacy. The probability that the person has the disease depends also on the *overall prevalence* of the disease in the population. If, for instance, only 1 in 1 000 000 in the population actually have the disease, even if the test is 99 % accurate, it will give vastly more false positive results than true positive ones. Thus, in this kind of context, with any given positive test result, it is overwhelmingly more likely that the person does *not* have the disease.

[42] Of course, for Worrall (2006), the NMA is invoked with reference to *logical* predictive successes.

With regard to the NMA, the situation is, apparently, analogous. The realist believes that if a theory passes a certain test, e.g. 'surprising novel predictive success,' the probability that the theory is true is very high. Similarly, she believes that the probability that a false theory would generate such success is very low. Based on this, she infers that a novelly successful theory is very likely to be true. However, the problem is that the realist is now ignoring the base rate of true theories within the *overall population* of theories. If the overall prevalence of true theories is very low, even if surprising novel success is a reliable test for truth, it can produce many more false positive results than true positive results. The realist is then faced with a problem. One the one hand, if she ignores the relevant base rate, the NMA becomes formally invalid (see Howson 2013). One the other hand, if she adopts a suitable prior belief about the relevant base rate, she will already have presupposed that the realist position is correct. In this, she will be unable to sway the anti-realist, who may hold different assumptions about the relevant base rate.

Realists have responded to this objection in multiple ways. One response is to challenge the very idea that a probabilistic reformulation of the NMA in terms of base rates makes sense in the context of scientific theorizing (see Worrall 2005; Psillos 2006). Scientific theories arise in very different contexts, they face very different competition, and they have very different features and properties. There appears to be no straightforward way of circumscribing what the relevant "population" of theories should be in any given case. For example, it surely does not make sense to claim that the probability that general relativity contains true theoretical constituents depends on the base rate of true theories in the social sciences. However, even a comparison to other theories in physics may not be appropriate. It still seems relevant to take into account that theories in physics have different types of virtues and more or less evidence behind them; for example, general relativity is supported by a much larger pool of evidence than many other theories, and it has stood the test of time for longer than many other theories. What is the relevant overall population to which this theory then belongs (e.g. theories supported by a pool of evidence of size x, theories held by scientists for x years)? The realist holds that there is no straightforward way of referring to base rates in this kind of a heterogeneous context where theoretical properties and the evidence for theories vary greatly from one situation to another (see, for example, Psillos 2006, p. 148-149).

A similar point against the base rate objection has been raised more recently by Henderson (2017), who argues that this objection fails to defeat the NMA because it relies on an assumption about the random sampling of individuals from a population. This assumption, she argues, is clearly not met in the case of science – at least not from the realist point of view. Rather, to be a scientific realist is to believe precisely that scientists select theories in a way that is not random but truth-conducive. She holds that a commitment

to a non-random selection process for theories can then back up particular Miracle-arguments about the truth of individual theories.[43]

A further counterargument against the base rate objection arises from recognizing the abductive nature of the NMA (see Worrall 2005). The central point of the No Miracles argument, as we saw in the previous section, is that surprising novel success calls for some *explanation*. Worrall (2005) argues that the Bayesian reformulation of the argument is not successful in fully formalizing the explanatory import that the abductive inference brings to the table. Rather, he maintains that any attempts to formalize the NMA in probabilistic terms are ultimately grounded on pre-theoretical intuitions. And, for him, these pre-theoretical intuitions set some form of realism as the default position with regard to the impressive predictive successes of science.

In my view, the base rate fallacy objection is suggestive, but it alone is not enough to defeat the realist position. The realist can respond that the process of scientific theory selection creates a biased rather than random sample of theories, and so the base rate objection does not apply (see Henderson 2017). Furthermore, theories have enough relevant differences to raise doubts about whether there is any coherent way to assign priors to the probability of truth; the best way to set up and present the NMA appears to be the abductive one (see Psillos 2006). Finally, the NMA is also motivated in an important way by intuition (see Worrall 2005, 2011; Chakravartty 2017a). In this, it is not alone among philosophical arguments. In fact, the intuitive argument to the opposite conclusion about science, i.e. the PMI, is also subject to a similar formal challenge (see section 3.3). In this kind of a context, the realist appears no less rational than the anti-realist in choosing her preferred framework (i.e. realism).

Another important objection to the NMA has been presented by Frost-Arnold (2010). Frost-Arnold argues in an interesting way that the NMA fails based on the very scientific standard it advocates: scientific theories should be accepted if they achieve novel predictive success. No such success, of course, has been enjoyed by the NMA itself. More precisely, the NMA

[43] Dawid and Hartmann (2018) have recently suggested another interesting way to defend the novelty-based NMA against the base rate objection. They argue that the NMA should be reformulated to concern the frequency of emergence of novelly successful theories in a particular scientific discipline or field. Realism is then defended by arguing that novelly successful theories in a particular field emerge at a high enough rate to produce a warrant for realist commitment. Roughly, the idea is that scientists only need to sample a relatively small number of theories to arrive at novelly successful theories, which indicates that the process is truth-conducive. There are some challenges to this argument, however. For example, we need some appropriate grounds for belief that true but successful theories are more common than false but successful theories in a particular field (see Dawid & Hartmann 2018). It also seems to me that there is something unsatisfying about hinging one's realist commitment to particular theories on the successes or failures of other theories. For example, what should we believe if scientists in some field generate and reject a thousand false theories, but there is one theory that clearly has superior (novel) successes over all the others? Should this theory not be rewarded based on its individual success, rather than let the failures of other theories stand in the way (cf. Worrall 2005)? If possible to defend, it appears that a theory-based realism would be preferable, although the frequency-based defense remains an alternative that warrants further investigation. See also Boge's (2020) criticism of Dawid & Hartmann (2018).

becomes unacceptable if one advocates a naturalistic philosophy of science, where philosophers should employ no other methods than those used by scientists themselves. Given that the NMA does not meet scientific standards, Frost-Arnold argues that naturalist philosophers of science should not accept it.

Here, the realist has at least two responses, both of which are acknowledged by Frost-Arnold. One is to hold that scientific realism itself is a philosophical rather than a scientific thesis. Worrall (2005), for example, considers and accepts the point that the NMA does not make any novel predictions, but argues that for this reason scientific realism should be seen as a philosophical rather than a scientific position. Novel predictive success is not a standard that philosophical theories are generally subject to, so scientific realism could be successful by *philosophical* even if not scientific standards. Likewise, Saatsi (2018) argues that the naturalistic approach is broad enough to incorporate philosophical statements that themselves do not enjoy the same degree of support that scientific statements do. In other words, in so far as the realist does not endorse a strict naturalistic position, the NMA remains available as a philosophical argument. Another response that is available to the realist is to develop a version of the NMA that does achieve novel predictive success. This, in fact, is a strategy that has been outlined in the context of the selective realism debate (see Vickers 2013; 2017). If the realist is able to produce a selective realist criterion that enables novel predictions about which scientific posits will be abandoned in the future, the NMA itself could become novelly successful. This strategy could ultimately remove the threat of self-refutation even for a strict naturalist.

Given the many options realists have available in responding to Frost-Arnold (2010), I do not believe that the self-refutation argument is enough to defeat the NMA. The argument stands at least by philosophical standards. However, as far as the selective realist strategy does not ultimately generate novel success for scientific realism, the self-refutation argument does reveal a certain *weakness* in the novelty-based defense of scientific realism. Even if novel success is not *necessary* for justified belief (as contemporary predictivists believe), its lack does indicate that one's belief should not be as strong (per predictivist criteria). For this reason, we may take Frost-Arnold's argument as one that is successful in undermining the novelty-based defense of scientific realism to some extent.

Other arguments for (substantive) scientific realism. Before we move to discuss the PMI, it is useful to consider briefly other possible strategies of defending scientific realism, and their relationship to the novelty-based NMA. This enables us to better understand what specific value the NMA has for the scientific realist (and what would be lost if the NMA fails), and help us contextualize the role of the NMA within the realist position.

Doppelt (e.g. 2005, 2014) has recently argued for a new IBE-based realism, which he calls 'Best Theory Realism' (BTR). Doppelt (2005) rejects the novelty criterion of success by arguing that a) theory confirmation

cannot depend on contingent facts about when the evidence was discovered (cf. Chapters 1 and 5), b) scientific realism itself fails on the novelty criterion (cf. Frost-Arnold 2010), and c) scientific realism should not explain only why theories successfully predict some data but also why they achieve explanatory success. He argues that a viable form of realism needs to be able to establish that the best explanation for the explanatory success of our best theories is their truth, and also simultaneously explain why the explanatory success of past theories need not mean that they are true. BTR holds that this is possible with our current best theories, i.e. we can retain realist commitment to them because they provide the best explanation for their own as well as their predecessors' success. BTR thus proposes that there is radical discontinuity between previous and current best theories. Some of our current best theories have far greater explanatory and predictive success than their predecessors, and they also have the explanatory capacity to explain the success of their predecessors. For this reason, we are justified in committing to them by IBE. The PMI, in turn, is defeated, because given the discontinuity between past and current best theories, past failures do not undermine our current best theories.

Fahrbach (2011a; 2011b; 2017) argues in similar fashion that many of our current theories enjoy far greater empirical success than previous theories. He points out that the amount, diversity, and precision of scientific evidence has increased enormously in the recent past, which has contributed strongly to the predictive success of many current theories.[44] Our current theories have been tested with much more diverse sets of evidence to much greater detail than previous theories, and yet they have *not* been refuted. Fahrbach argues that the growth of scientific evidence creates an important discontinuity with the past, and for this reason the realist can resist the PMI: current theories are not as likely to fail as past theories were. Fahrbach (2017, p. 5041) emphasizes that his purpose is to formulate one path forward for the realist, rather than reject others. He continues to rely partly on the NMA to make the case that the success of many current theories is enough for realist commitment, arguing that the kind of success enjoyed by current theories is high enough to make the NMA "extremely plausible" (ibid. p. 5068).

An important problem in Doppelt's and Fahrbach's proposals is how we justify the assumption that the present is relevantly different to the past in science. Wray (2013) argues against Fahrbach that his new argument about the growth of evidence does not establish a relevant discontinuity between the past and the present: the same exact argumentation strategy would have been available to past scientists with reference to their theories and sources of evidence. Yet, as history shows, they would have been wrong to make such claims. Thus, the PMI remains intact: a similar fate may yet await us and our theories. Similar considerations can be advanced against Doppelt's

[44] Fahrbach does not appeal to novel predictive success, but understands 'predictions' simply as the observable consequences of scientific theories that scientists are aware of (see Fahrbach 2011a, p. 1286).

arguments (see Alai 2017; Díez 2018). Alai (2017, pp. 3283-3284), for example, points out that the growth of evidential standards in science generates a particularly strong PMI against Doppelt's proposal: as evidential standards continue to improve, we might expect that even better theories than our current ones emerge in the future that are able to meet these standards, leading to the abandonment of some current theories. But, in so far as this happens, our best current theories could after all be *completely false* by Doppelt's criteria. Alai argues that the realist is better served by sticking with selective realism. Selective realism accepts that something even better may yet come along, but maintains that we can still be justified in believing in many constituents of currently successful theories (via the NMA).

Another strategy in defending scientific realism appeals to the notion of approximate truth, or *truthlikeness* (see Popper 1963). Certainly, just about all realists appeal in some sense to notions of approximate truth, because they believe that scientific theories need not be completely true to be able to latch on to some parts of reality (see, for example, Psillos 1999). There is also an intuitively clear way in which theories can be close to truth without hitting on it exactly and literally; for example, a prediction from a theory may deviate marginally from measurements, and thus only be 'approximately' rather than literally true, but it could still provide strong grounds for realist commitment. Some realists hold that the idea of approximate truth is intuitively clear enough without more detailed formal analysis, and as such supplements other realist arguments (e.g. Psillos 1999). However, it is also possible to ground one's realism completely on the notion of truthlikeness. This is the strategy of Niiniluoto (2017), who argues that science progresses in terms of truthlikeness, where new theories that progress empirically beyond older theories should be considered more truthlike than their predecessors. In this approach, we never know for sure how close to the truth we are, but realism becomes viable in a comparative sense, where science can be seen as gradually progressing closer and closer to the truth.

Comparing these other strategies of defending scientific realism to the novelty-based NMA, a natural option is to view them as *complementary* rather than competing strategies (see Fahrbach 2017, p. 5041; Chakravartty 2017b, pp. 3388-3389). The growth of scientific evidence, and the impressive success of some of our current theories, could provide some further grounds for realist optimism about the theoretical constituents of current best theories. Yet, we need not commit to these theories completely, as the selective realist recommends: even current best theories could contain some idle constituents that will be abandoned in the future. The truthlikeness approach and novelty-based selective realism could likewise be both accepted by realists (see Chakravartty ibid.). The scientific realist could accept that science achieves progress in terms of truthlikeness, but also be interested in using novel predictive success to identify particular working and idle posits in scientific theories in looking for more substantial commitment to truth.

In all of this, the role of the novelty-based NMA is quite interesting: it is the one argument available to the realist that enables her to identify particular theories, or particular constituents of theories, that warrant realist commitment, without being subject to the charge of ad hocness with regard to our current historical period (contrary, for example, to BTR). That is to say, the novelty-based NMA can be applied to all theories – past, present, and future – as a criterion for what warrants realist commitment. If supplemented with an appropriate selective realist criterion, it provides an answer to the question of which scientific posits or parts of science should be thought of as (more) likely to be true. This may be one of the reasons for why the NMA has been considered so central to the defense of scientific realism. It is the one argument that enables the realists to put their foot down and throw their commitment behind particular theories and theoretical constituents. If the novelty-based NMA were to fail, scientific realists would apparently have to retreat to much more modest forms of realism, where they may only be able to hold that we are approaching the truth without ever knowing how close we have got (e.g. Niiniluoto 2017), or face a very strong PMI, which threatens even our best current theories (see Alai 2017). I will continue to explore the NMA, and its viability in the defense of scientific realism, at different points in subsequent chapters.

3.3 THE PESSIMISTIC META-INDUCTION ARGUMENT

In contrast to scientific realism and the NMA, scientific anti-realism is motivated by another intuitive argument about scientific practice: the Pessimistic Meta-Induction (PMI) argument. The anti-realist points to episodes of theory change in the history of science, and argues that growing empirical success in science need not correspond to advances in theoretical truth: the past successes of theories that are now known to be false show that success can follow from false theoretical constituents in science. Given the historical record, we should conclude that the same fate might yet await our current best theories, and thus refrain from committing to their truth.

The PMI has gone through a period of coevolution with the NMA. After the realists specified that only novel successes count, anti-realists shifted their strategy to highlight novelly successful but false theories (e.g. Lyons 2002; 2006). After the specification that the novel predictions should be surprising or impressive, counterexamples have emphasized the surprising and impressive novel predictions of false theories (e.g. Carman & Díez 2015; Tulodziecki 2017; Rossetter 2018). Similarly, with the introduction of the selective realist strategy, there has been an attempt to demonstrate that the impressive novel successes of abandoned theories really did derive from false theoretical posits. We shall next examine more closely how the PMI is meant to defeat the realist position.

The form of the argument. Similarly to the NMA, there has been debate about how the PMI should be formulated. The most straightforward formulation is a simple inductive one (see Chakravartty 2017a). From the present point of view, most past theories are false. In fact, we can see that this holds from the point of view of most times. Hence, by enumerative induction, there is likely to be a future time when our current theories are considered false. Accordingly, our current theories are false. (With the introduction of the more stringent criterion of surprising novel success, we can change this to: from the present point of view, many theoretical posits that were deployed in the past in the derivation of surprising novel successes are false. This holds for most times. Hence, this is also likely to be the case with regard to the theoretical posits of our current novelly successful theories.)

In other versions, the PMI is formulated as a deductive argument. Lewis (2001, p. 373; see also Psillos 1999, pp. 97-98), for example, reconstructs the argument as a *reductio*:

(1) The success of a theory is a reliable test for its truth.
(2) Most current theories are successful and hence true.
(3) Most past theories are false, since they differ from current theories in relevant ways.
(4) Many of the past false theories were successful.
(5) Success is not a reliable indicator of truth.

Lyons (2002; 2006), in contrast, argues that the PMI should be understood as a 'Pessimistic Meta-Modus Tollens:'

(1) If realism is correct, then each theoretical constituents that were deployed in the derivation of key successes are true.
(2) There are examples of false theoretical constituents that were deployed in the derivation of key successes.
Therefore, (3) the realist hypothesis is false.

Concerning the deduction-based alternatives, they appear excessively strong to represent what the realist is committed to (see Vickers 2013, 2017). Based on Lyons's argument, just one false theoretical constituent that played a role in the derivation of a novel success is enough to defeat realism. Yet, hardly any realist would wish to commit to such a strong link between novel success and truth (cf. Vickers ibid.). This, in fact, already follows from the basic nature of the NMA: as we saw in the previous section, the NMA is presented as an abductive argument, and as such it is not meant to guarantee truth. The realists are looking for criteria of success that permit a *fallible* inference from novel success to the truth of key theoretical constituents. This does not require a commitment to a deductive, truth-preserving argument.

With the acknowledgement that the realist is not after infallible indicators of truth, the discussion generally turns to how *strong* of an

inductive or statistical argument the anti-realist can raise against the realist (see Wray 2015, p. 64). From this point of view, even one clear example will be problematic to a realist who holds that novel success is a *very* good indicator of truth (see Vickers 2017, p. 3223). If there are several examples, the PMI starts to threaten even more modest realists who think that novel success makes truth at least probable. As a general rule, the more examples there are, and the more heterogeneous contexts they are found in, the stronger the PMI becomes.

If interpreted as an inductive or statistical argument, the PMI faces similar formal challenges as the NMA (see Lewis 2001; Magnus & Callender 2004). An advocate of the PMI points to a particular historical period, observes that the ratio of (what are taken to be) true and successful theories to false and successful theories is small by current lights, and then projects this to concern also the present time. However, in this she presumes that the relevant population of theories remains constant throughout history; i.e. that the relevant ratio remains roughly the same as science progresses (see Lewis 2001). Yet, similarly as with the NMA, this requires the anti-realist to make assumptions about the relevant base rates, of which independent evaluation is not possible. The assumption that there is a constant base rate is precisely what the realist would question (see Henderson 2017). This seems to leave the anti-realist incapable of defeating the realist position.

The PMI can be defended against this charge in a similar way as the NMA. We can ask whether it makes sense to formalize the argument in terms of statistical probabilities (cf. section 3.2). Saatsi (2005b) argues that the PMI is simply concerned with *severing* the connection between the realist's chosen criterion of success (e.g. novelty) and truth. The intuitive problem that the PMI raises for the realist is that (novel) success *could* also arise from radically false theoretical constituents. In other words, the primary force behind the PMI is that it acts as *a reductio* of *the abductive* NMA: apparently, (novel) success arising from falsehoods is not that miraculous after all. Again, a single example will not be compelling, but a few will be enough to raise a *prima facie* challenge, and a large and heterogeneous sample can potentially wreak havoc on the realist's NMA: the realist is forced to concede that truth is not that good of an explanation for novel success after all.

We should next examine a few case examples that the anti-realists have appealed to in building their case for the PMI.

Recent examples. The discussion on potential counterexamples to the realist position has been very active for the past two decades. Vickers (2013) provides a list of twenty potential examples where novel successes apparently followed from false theories (the debate is ongoing without generally accepted conclusions in many cases). These include the caloric theory of heat, Fresnel's theory of light and the ether, Rankine's vortex theory of thermodynamics, Kekulé's theory of the Benzene molecule, Bohr's prediction of the positron, the phlogiston theory, etc. The selective realist studies such cases, and seeks to show that any novel predictions in these cases

did not really depend on false theoretical constituents (see, for example, Psillos 1999). An opponent of selective realism attempts to do just the opposite, i.e. show that false constituents were deployed in the derivation of the novel successes (e.g. Lyons 2002; 2006).

For the purposes of this study, I want to introduce in particular a few more recent examples that emerged after Vickers's (2013) collection of examples. These examples are particularly relevant to the present discussion in that the authors meet all of the realist's criteria that were specified in the previous section. The authors emphasize that the cases involve *surprising* novel successes, and yet the relevant *theoretical constituents* appear wildly false. These examples are particularly well suited to undermine the realist's claim that truth is the best, or the only, explanation for impressive novel success in science.

1) *Ptolemaic astronomy*. Carman and Díez (2015) provide a detailed case study of theoretical constituents in Ptolemaic astronomy. They proceed to show that the false postulates, (i) Mercury and Venus circle the Earth closer than the Sun and (ii) the orbits of the planets cannot overlap, were necessary in deriving the novel predictions that (a) only Venus and Mercury pass in front of the Sun, (b) outer planets beyond the Sun are never eclipsed by the Earth shadow, and (c) the phases of the outer planets proceed from full to almost-full and then back to full. The assumptions i and ii are central postulates of Ptolemaic astronomy, they are clearly false, and yet they generated multiple novel predictive successes for Ptolemaic astronomy. Carman and Díez (ibid. p. 29) argue that some of the predictions are as surprising (e.g. c) as the predictions of Heliocentrism that were later used to refute Ptolemaic geocentrism.

2) *The zymotic theory of disease*. Tulodziecki (2017) presents a case study of the 19th century miasma-based theory of disease, according to which diseases are caused by the interactions between miasmas and zymotic materials. Tulodziecki identifies three core working posits of the zymotic theory: (i) organic matter given off by decaying material is the cause of disease, (ii) this organic matter is suspended in the air, and (iii) it is transmitted through the air. None of these posits are preserved in the modern germ theory of disease. Yet, Tulodziecki shows that they were deployed in the derivation of numerous novel predictions about the geographical incidence of disease, the relationship between air quality and disease, etc. Most pertinently, Tulodziecki finds an elevation law relating cholera mortality to soil elevation that was derived based on the zymotic theory. This law was subjected to multiple use-novel tests at the detailed level of sub-districts in London and other cities, and on each occasion, the relationship between cholera mortality and soil elevation was found to obey the (spurious) law.

3) *Hutton's theory of the Earth*. Rossetter (2018) analyses Hutton's theory of the Earth, according to which the Earth is an organic body that possesses a reproductive mechanism to repair eroded matter. Rossetter divides Hutton's theory into eight constituents, and argues that while three of

these constituents are approximately true by modern lights, at least three are clearly false (and two more are questionable) – including the central posit about the reproductive mechanism. He argues that the false and questionable posits were deployed in the derivation of two very specific predictions concerning granitic veins and inclined layers of older and newer strata. These predictions were impressive enough to win over skeptical converts in the scientific community at the time. Yet, the theoretical mechanism that Hutton postulates is wildly wrong from the modern perspective.

These examples appear particularly difficult for the contemporary selective realist. The authors take pains to pare down the theoretical posits that are needed to derive the relevant novel predictive successes, responding directly to criteria set in recent selective realist accounts (e.g. Vickers 2013). There does not appear to be any obvious way to further eliminate or weaken the posits in question and still retain the predictive successes. For example, in the case of the zymotic theory of disease, the posits are effectively weakened to state that "*whatever* is given off during decomposition is transmitted through the air and involved in causing diseases" (Tulodziecki 2017, p. 1005). Anything less removes the bare minimum posits of decomposing sources and air as a medium that are necessary to derive the successful predictions.

As far as the realist cannot find a way to account for these examples, they present a *prima facie* challenge to the abductive NMA. The realist believes that surprising novel success in the absence of truth would be miraculous in science; yet, we have apparently witnessed many miracles here. What is particularly interesting about these examples is that they show that multiple surprising novel predictions can be derived even from theories that differ greatly from modern theories. This may provide some preliminary indications that novel predictive success is not what the realists have assumed, although the precise implications of this finding remain unclear at this point in our investigation, as the evidence base remains relatively small.

Other anti-realist arguments. Before we move to estimate the overall strength of the PMI against the NMA, it is useful to briefly introduce two other important arguments for anti-realism: the argument from underdetermination and the argument from unconceived alternatives. Both of these present similar concerns as the PMI about there being (false) alternatives to current empirically successful theories, but they employ a different argumentative strategy.

The argument from underdetermination is one of the most famous in contemporary philosophy. Simply put, this argument states that when we are confronted with an empirical challenge to our theories or beliefs, what we learn is that *some* part of our belief system (or theory) needs to change in order to accommodate that challenge (see Duhem 1914; Quine 1951). *Which* part should change is underdetermined by the empirical evidence as such. There are multiple ways to alter what we postulate about the world to bring the empirical evidence back into agreement with our beliefs and theories. When applied particularly to theory change in science, the

underdetermination argument alerts us to consider that for any theory we postulate, there are many more that preserve the same empirical predictions while nonetheless altering some (or all) parts of the theory itself. With the many alternatives available, we are apparently unjustified in believing that the particular theory we have postulated is the true one.

A chief problem with the underdetermination argument, raised forcefully by Laudan (1990) (and extended by Laudan & Leplin [1991]), is that the fact that there are *logically* equivalent alternatives to our beliefs and theories does not mean that these alternatives are equally *rational*. Laudan and Leplin (1991) argue that scientific theories can enjoy different degrees of support from how they relate to other things we believe and have epistemic support for, over and above their specific empirical consequences (e.g. other theories, well-confirmed scientific principles). It is not clear that there really are many equally rational alternatives to our theories – or at least whether this the case should be evaluated on a case by case basis rather than assumed by fiat. In accordance, many realists have remained unpersuaded by the mere possibility of alternatives. Kitcher (1993, p. 154) puts it succinctly: "give us a rival explanation, and we'll consider whether it is sufficiently serious to threaten our confidence."

The debate on the underdetermination argument, particularly in its general form, spans well beyond this study. However, a few words need to be said about its relationship to the NMA and the PMI. First, the contemporary version of the novelty-based NMA appears particularly well positioned to answer to the underdetermination challenge. The contemporary realist grants that theory confirmation can be underdetermined by the empirical evidence as such: there are many ways to theorize about the same empirical evidence. However, the realist argues that there is special type of evidence that speaks to the truth of certain theoretical constituents: novel predictive success. This is precisely the kind of success that cannot be achieved after-the-fact by producing empirically equivalent alternatives with the specific purpose of accommodating for the predicted results. In this way, novel predictions can reveal actual true constituents even if there are many logically equivalent ways to accommodate particular empirical results. Second, the PMI, in turn, pushes back against the realist's NMA, and helps bolster the underdetermination challenge. The PMI shows that novelly successful but false rival theories are not a mere possibility but something that have actually emerged in scientific practice. In other words, the PMI helps re-establish that underdetermination poses an actual challenge even in the context of scientific practice, where many constraints on rational theory confirmation are already accepted by scientists.

Another important anti-realist argument is *the argument from unconceived alternatives*, raised by Stanford (2006). Stanford argues that the standard argument from underdetermination has failed to provide a compelling general case for that there are empirical equivalents to all or most of our scientific theories. Instead, he argues that there is another issue that arises from the history of scientific practice itself: competing theories that are

not empirically equivalent but are *equally well confirmed* at the time. In the scientific process, there has been a reoccurrence of situations where a theory that was held at a particular time was challenged by a new theory that was unconceived of before. When the new theory emerges, it is shown that there is an equally well confirmed (and, ultimately, better) alternative to the old theory. Given the myriad of times this has happened in the history of science, Stanford argues that the same fate may yet await our current best theories. In other words, our current theories are threatened by equally well confirmed alternatives that scientists have simply not conceived of yet.[45]

The selective, novelty-based defender of scientific realism appears, again, well positioned to answer to this challenge. The novelty-based selective realist can freely grant that there may be yet many important, unconceived scientific theories. In fact, she believes firmly that science is going to progress in the future as well. New theories may very well emerge that enable scientists to leave behind some current idle theoretical posits. However, the novelty-based selective realist holds that this does not mean that many current theoretical constituents are false. On the contrary, based on impressive novel predictive success, we can recognize that our current theories contain many important true constituents. Conversely, for an anti-realist who appeals to the PMI, the argument from unconceived alternatives could potentially function as an extension of that strategy. The anti-realist could look for many different ways in which something that was unconceived of for a time within the scientific community later led to revisions that forced scientists to abandon theoretical posits associated with novel predictive successes. This could reveal some further similarities with the present situation, providing more fuel to the anti-realist case.

The PMI vs. the NMA. Where does the debate between the realists who appeal to the NMA, and the anti-realists who appeal to the PMI, stand? Vickers (2018, p. 118), writing from the realist perspective, holds that "the selective realist is in a strong position vis-à-vis the historical challenge." Stanford (2020), writing from the anti-realist perspective, argues that the burden of proof has now shifted to the realist. Others have argued that the debate has come to a stalemate, where neither position can claim to be the more rational philosophy of science (see Forbes 2017).

In my view, it is not perhaps accurate to say that the debate has reached a stalemate, as advances continue to be made on both sides. However, the discussion is certainly *limited* by the nature of the respective arguments both for and against each position. Both the selective realist and the historically motivated anti-realist depend heavily on case studies. The viability of the selective realist position depends on whether the realist is able to defend the claim that theory changes have not led to the abandonment of novelly

[45] Rowbottom (2019a, pp. 62-77) expands Stanford's argument, suggesting that it should be applied even more broadly. He argues that the problem for the realist is not just unconceived theories, but also unconceived observations, predictions, explanations, models, experiments, methods, instruments, and values.

successful theoretical posits. The PMI-based challenge depends on providing the opposite kind of examples. In both cases, the analysis of the examples requires a great deal of effort, limiting the number of cases that can be afforded serious discussion. Many more relevant cases are bound to exist in the scientific literature than those that have been discussed so far. This raises concerns about how much can be inferred based on such small samples. We may be very far from a situation where there is enough evidence to uncontroversially support one position over the other.[46] In the meantime, both positions continue to be motivated by compelling intuitions about what best explains central features of scientific practice: its impressive (novel) success on the one hand (NMA) and major episodes of theory-change on the other (PMI).

Where does this leave us? In my view, the novelty-based NMA remains standing, but the anti-realists have been successful in raising a *prima facie* challenge to it. There are enough examples from the history of science to suggest that novel predictive success may not be as powerful as realists may sometimes have presumed (e.g. Carman and Díez 2015; Tulodzieki 2017; Rossetter 2018; see also Lyons 2002, 2006). This is not enough to defeat the NMA, as the number of examples remains relatively small. However, this suggests that there may be problems that have so far not been fully appreciated. Further clarity on the issue could be reached in a number of different ways. One strategy is to gather much more evidence about abandoned and (so far) retained theoretical posits (i.e. continue business as usual). Another strategy is to develop new arguments that challenge either position on more direct terms. Finally, the entire premise of the current debate could be challenged, and realism or anti-realism (or, perhaps, some other position that rejects both) defended in some entirely different way. We will return to this issue in Chapter 5, exploring in particular *the second strategy*.

3.4 CONCLUSION

We have now arrived at a place where we are able to draw conclusions about where the discussion about the epistemic role and value of novel prediction in science stands. There are two ways to defend the view that novel predictions are special:

A) *The negative thesis*. Novel prediction provides stronger confirmation to scientific theories than accommodation because accommodation is associated with negative epistemic consequences. Accommodation may lead to overfitting (Hitchcock & Sober 2004), fudging (Lipton 2004), or questionable hypothesis hunting (Mayo 1996). These issues

[46] In an aptly titled paper, "Historical Inductions: New Cherries, Same Old Cherry-Picking" (2015), Mizrahi argues that advocates of the PMI are guilty of cherry-picking evidence in support of their position. The same could be alleged about the supporters of selective realism.

do not concern novel prediction (or, at least, not to the same extent as accommodation), so novel prediction has an advantage.

B) *The positive thesis*. Whereas accommodative success can be explained by the theorist's intention of accommodating the evidence, novel predictive success calls for some other explanation. The best, if not the only, explanation for novel success is that the theory contains true constituents that were deployed in the derivation of that success (or, that scientific methodology was reliable in producing such a theory). In virtue of this explanatory advantage, novel success provides the grounds for realist commitment to the truth of working constituents of scientific theories.

The two predictivist theses, even if supportive of one another, are independent, and they present different types of argument for the special advantage of novel prediction. The negative thesis rests on a primarily causal argument, where accommodation is thought to cause (or correlate with something that causes) for the theory selection, testing, or generation process to become more unreliable. The positive thesis is based on an explanatory argument, where truth or reliability is invoked as the best explanation for novel predictive success. One could endorse both, one or the other, or neither argument. The negative thesis is more modest: it holds merely that novel prediction is "better" than accommodation. This need not necessarily mean that novel success makes truth likely – just more likely than in the case of accommodation. The positive thesis extends the predictivist position further by arguing that novel success specifically speaks to theoretical truth.

In order to evaluate whether or not there are special advantages to novel prediction, both the negative and the positive thesis need to be further investigated. If either thesis holds, predictivism (of a certain type) holds. If both fail, predictivism is shown inadequate. This evaluation is conducted in the next two chapters.

4 THE NEGATIVE THESIS: THE PROBLEM(S) OF 'BAD' ACCOMMODATION

The negative predictivist thesis holds that novel prediction provides stronger confirmation than accommodation because accommodation is associated with negative epistemic consequences. In the case of accommodation, there is a chance that various kinds of 'bad' accommodation may have taken place, so the information that evidence was accommodated rather than novelly predicted should prompt evaluators to lower their confidence in the theory. In the case of novel predictive success, there is evidence that bad accommodation has not taken place, so the theory receives unrevised support from the evidence (i.e. theory confirmation depends on the contents of the theory, the evidence, and background assumptions as such).

This chapter evaluates the negative predictivist thesis. Is there an important confirmatory asymmetry between novel prediction and accommodation in science due to the problem(s) of bad accommodation? Before we attempt to evaluate this claim, a few more words need to be said about its constituent parts. What does it take for there to be (i) *an important confirmatory asymmetry*, (ii) *in science*, (iii) *due to* the difference between novel prediction and bad accommodation?

First, a confirmatory asymmetry that is due to the distinction between novel prediction and accommodation means that the information that some evidence was accommodated rather than novelly predicted carries epistemic import. The evaluation of a scientific theory requires that a number of factors be taken into consideration. For one, there is the evidence as such. For example, if your hypothesis is "all crows are black," you will be interested in the contents of "crow-observations" as such. The same holds for just about any context of theory evaluation. For another, there is the theory as such. Theories have different features and properties that may be epistemically relevant (see Keas 2018 for a review). One example is simplicity. If your hypothesis is "all crows are black except this particular one today at this particular location," you appear to be adding extra complexity to your hypothesis that has costs in terms of epistemic support. Finally, there is background knowledge or information. This covers general knowledge or principles that are already accepted or taken for granted in a particular context of theory evaluation (see, for example, Worrall 2009). Examples might include other theories that have already been accepted in the scientific community, fundamental scientific principles (e.g. the conservation of momentum), overall background assumptions (e.g. "inductive successes are not due to a hoax by a grand skeptical conspiracy"), etc. A confirmatory asymmetry that is due to prediction and accommodation means that after taking into account the

contents of the theory, the evidence, and background knowledge, the fact that some evidence was novelly predicted rather than accommodated still carries epistemic import.

Second, in order for the prediction versus accommodation distinction to be in any sense important, novel prediction must matter due to its own particular virtues in relation to (bad) accommodation; that is, an epistemic asymmetry in scenarios where we compare novel prediction and accommodation should be attributable particularly to this distinction rather than some other, more general epistemic principle. There are certain trivial ways in which novel predictions are associated with epistemic advantages (cf. section 2.1). For example, in case of temporal novelty, we are always adding more new evidence for the theory, which is obviously better than having less evidence. Anti-predictivists such as Harker (2008) argue that the distinction between prediction and accommodation ultimately reduces to these other epistemic considerations, e.g. a preference for progress. If the prediction versus accommodation distinction is to count as such, there must be a difference that arises particularly because of the benefits that novel prediction provides in relation to accommodation. So, for example, in the case of temporally new evidence, the negative predictivist thesis holds that the advantage of novel prediction goes beyond just the benefit of adding more evidence. The fact that the evidence was not used in the construction of the theory brings *more advantages* because it provides counterevidence about bad accommodation.

The importance of novel prediction requires further investigation also in another sense, namely, in terms of its relative strength. It is one thing for there to be some contexts in science where novel predictions count for more. It is another question if novel prediction is the best way to combat bad accommodation in science, and yet another question how much stronger confirmation novel prediction provides over accommodation. Even anti-predictivists agree that some forms of weak predictivism may come into effect in some situations (see, for example, Harker 2008). However, the advantage of novel prediction could turn out to be entirely trivial, perhaps because it is so weak, or because there is something better available than novel prediction. For this reason, we will pay close attention to the relative advantages of novel prediction in this chapter. An important goal is to evaluate what value novel prediction carries overall among other epistemic considerations in science.

Third, the problem at hand concerns scientific practice. We are interested in the question, should novel prediction provide an epistemic advantage over accommodation in science? The issue is evaluated from the standpoint of a scientific evaluator who examines the theory and the evidence, and attempts to come up with an estimate of the degree to which the evidence supports the theory. In order for the negative predictivist thesis to hold – either in a general sense (e.g. Lipton 2004) or at least in some important scientific contexts (e.g. Mayo 1996) – the prediction/accommodation distinction must be epistemically relevant to scientific evaluators. There are

also other perspectives to theory evaluation, e.g. that of the layperson (see Barnes 2008). The results of this investigation will ultimately concern also the layperson, but this is a point which we will not consider until the end of Chapter 6, after the issue of scientific confirmation has been fully evaluated.

The advocates of negative predictivist accounts argue that there is the kind of predictivist advantage in science that has been laid out above. The negative predictivists hold that accommodation can have negative epistemic consequences in scientific practice, and therefore the information that some evidence was novelly predicted rather than accommodated is (sometimes or often) relevant for scientific evaluators. The issue is not accommodation as such, but the problems that accommodation sometimes leads to (or is associated with). These include overfitting (Hitchcock & Sober 2004), fudging (Lipton 2004), and hypothesis hunting (Mayo 1996). Scientific evaluators may be unable to rule out these forms of bad accommodation directly, so novel prediction brings epistemic advantages.

I will now proceed to evaluate the negative predictivist thesis. In section 4.1, I clarify further what it means to consider this thesis in the context of scientific practice. I discuss certain intuitive examples that appear to show a predictivist advantage that could be attributed to problems of bad accommodation, but argue that these examples fail to represent scientific practice. In sections 4.2 and 4.3, I consider the problems of overfitting and hypothesis hunting. Both of these are, in fact, among a general category of issues scientists have called 'Questionable Research Practices' (QRPs). The overall relationship between QRPs and the prediction versus accommodation issue is evaluated in section 4.3. In section 4.4, I assess Lipton's fudging argument for predictivism. Section 4.5 introduces a special case from science that is similar to the intuitive examples discussed in section 4.1: epistemic opacity. The relevance of epistemic opacity to the negative predictivist case is clarified in this section. Finally, section 4.6 concludes the discussion on the negative predictivist thesis.

4.1 INTUITIVE EXAMPLES

Predictivist arguments have often been motivated with intuitive "real-life" examples where accommodation seems to have little or no confirmatory force in comparison to novel prediction (e.g. Maher 1988; Mayo 1996; White 2003; Barnes 2008). It is instructive to examine these examples to explore further what the prediction versus accommodation issue is about in the context of scientific practice.

Consider that you are looking to put your money into an investment fund and there are two options available (see Barnes 2008). One investment fund is run by a person who has a long track record of predictive success with the fluctuations of stock prices. The other investment fund is run by a person who claims to have a theory that explains why stock prices have

changed in the past. However, he does not have predictive success to show for himself yet. You are given no more information about the case, e.g. the contents of the theories used by the investors in either scenario. Just about everybody agrees that, with this information, it is rational for you to put your money into the fund with the track record of predictive success. The impressive predictive success, it appears, demonstrates that the person who runs the investment fund is very likely to have some genuine insight into what moves stock prices. The accommodative success of the other investor provides no such guarantee. Similar examples that seem to illustrate a distinction between prediction and accommodation have been constructed around other hard-to-predict outcomes such as series of coin tosses (Maher 1988), lottery results (White 2003, pp. 668-669), terrorist attacks (Mayo 1996, p. 251), etc. For example, in a lottery case discussed by White, we are introduced to an individual who claims that the lottery has been rigged in favor of a particular person, A. If this hypothesis is presented to us *prior* to the drawing of the lottery, which subsequently results in A winning, it appears to provide very strong support for the rigging hypothesis. If it is presented *after* the outcome is already known, it appears to count for very little. In each of these cases, it seems that predictive success with unpredictable outcomes is more compelling than after-the-fact accommodation of those same outcomes.

Each of these examples could be interpreted as illustrating the negative predictivist thesis in action. The problem in the accommodation scenarios appears to be that we, as evaluators, have no way of verifying what the accommodators are claiming. We are simply presented with a claim, and then are offered no reasons or evidence for that claim. As far as we know, the accommodators could have engaged in any form of bad accommodation. This need not involve the forms of bad accommodation that have been discussed by the negative weak predictivists; we are all quite familiar with some more mundane forms of "bad accommodation." One example is just the act of lying. Consider the investment fund example again. In the accommodation scenario, we encounter an investor who claims to have a successful theory about the movements of stock prices. One possible explanation for this claim is that the investor really has such a theory. Another explanation is that the investor has no such thing and is merely trying to persuade us to part with our money. The issue, in any case, is that in the accommodation scenario we are aware that it is quite possible that the accommodative "success" is due to any number of factors but the actual possession of a successful theory – minimally, there is no theory at all and the person is simply making things up. Hence, absent more information, we have (at least) preliminary reasons to be suspicious about the accommodator. With impressive novel predictive success, on the other hand, we can rule out just about any type of bad accommodation, and increase our confidence that we are dealing with a genuinely successful theory or hypothesis.

It is uncontroversial that in examples that have been set up in this fashion, there is an epistemic distinction between what are set up as the

prediction and the accommodation scenarios. As Worrall (2014, p. 58) puts it, in these type of situations, "we would surely all, unless being perverse," place more trust in the predictor than the accommodator. However, the issue from the point of view of predictivism is, what do these examples reveal about the prediction versus accommodation issue in scientific practice? Here, a number of commentators on different sides of the predictivism debate have agreed that the answer is very little (e.g. Howson & Franklin 1991; Harker 2006; Worrall 2014; see also Howson 1988; Collins 1994; White 2003, p. 656). This is because there are important disanalogies between these cases and the cases that are typical in science. Most importantly, what is lacking here is *information about the contents of the theory and the independent evidence* used by the predictor and the accommodator (see Howson & Franklin 1991; White 2003, p. 656; Worrall 2014, p. 58).[47] Examples such as the investment fund case or the lottery case are presented in a vacuum of relevant information: we are introduced to a person who claims to have a good theory or hypothesis, but are either given no information about the contents of the theory or the evidence that it is based on. This is quite unlike cases in science, where theorists are asked to subject their theory and evidence for public evaluation. Predictivists and anti-predictivists alike have agreed that a compelling, non-trivial predictivist argument should be able to establish that the predictivist advantage arises in cases where the theory and the evidence are public rather than hidden from evaluation.[48]

To illustrate briefly how the situation changes if the theory and the independent evidence are made public, consider as an example earthquake prediction. To keep things simple, let us say that we are evaluating a theory that holds that a certain geological event Z raises the probability of an imminent earthquake. Evidence for this theory comes from two sources. First, there is independent longitudinal evidence that Z correlates with earthquake occurrence in locations all over the world. Second, general, well-confirmed theories in geology indicate that a causal relationship between Z and earthquakes is plausible.[49] Now, compare two scenarios. In the first scenario, theorist A examines the well-confirmed theories in geology and postulates that Z raises the probability of an imminent earthquake. A then predicts that Z has a positive association with earthquakes, finding out later that the longitudinal

[47] Harker (2006) argues that another important disanalogy is that in these type of examples we are considering the prediction of ostensibly random events. He argues that scientific theories are typically not formulated about such random events, so the analogy fails to extend to scientific practice.

[48] In cases of withheld information about the theory and the evidence, the epistemic distinction in the prediction and the accommodation scenarios follows from any number of confirmation theories without the need to invoke specifically a predictivist advantage. Minimally, all we need to invoke is the principle that a scientific theory or claim should be subjected to an intersubjectively evaluable test, and any confirmation theory that emphasizes independent tests delivers the same result (e.g. Worrall 2014). The conclusion also appears to follow naturally from Mayo's (1996) confirmation theory, which requires severe tests without invoking specifically the prediction versus accommodation distinction.

[49] It should be noted that no theory of this kind exists at the moment. Earthquake prediction is incredibly difficult, and scientists have yet to demonstrate that prediction is possible (see, for example, Geller et al. 1997, p. 1617).

evidence confirms this. In the second scenario, theorist B first examines the longitudinal evidence, and uses it to generate the same theory about Z. B then later appeals to the same geological theories. B thus accommodates rather than predicts the longitudinal evidence. The question then becomes, is there an epistemic asymmetry between these two scenarios?

For all that has been said about bad accommodation, the answer appears to be 'no.' Unlike in the intuitive cases, the accommodation scenario here is clearly not a case where there is a possibility that the theorist has been lying about the existence of a theory or independent evidence: we know in this case what the theory and the evidence are. There are also no asymmetries due to overfitting or hypothesis hunting, because in both scenarios the theory is supported by the same exact (independent) longitudinal dataset. Finally, the theory is extremely simple (it presents a modest claim about a causal relationship between two events or phenomena), so there appears to be very little reason why fudging should be suspected in one case but not the other. In other words, now that we have the theory and the evidence on the table, the bad accommodation problem goes away – in this case at least.

The intuitive examples enable us to illustrate more clearly what the prediction versus accommodation problem is about in the context of scientific practice. The predictivists have argued that the prediction versus accommodation distinction is relevant in the kind of cases that feature in science. To evaluate this specific claim, we need to examine situations that are relevant to science. This is not the case – generally, at least – with artificial examples where relevant information is withheld from evaluators. We shall next look at the forms of bad accommodation that are meant to show an epistemic advantage of novel prediction for scientific evaluators. In section 4.5, we come back to the issue of missing or imperfect information and the negative predictivist thesis, considering it as a special case.

4.2 OVERFITTING

The first negative predictivist argument we examined in this study is that of Hitchcock and Sober (2004). Hitchcock and Sober argue that unless appropriate measures are taken to combat overfitting in statistics, accommodation risks *overfitting* the data – i.e. our model becomes too sensitive to noise in the dataset, hindering its out-of-sample predictive performance. Novel predictive success with out-of-sample data indicates that problematic overfitting has not occurred, and so it confirms the model more strongly.

The first thing to recognize about Hitchcock and Sober's weak predictivism is that it is *very* weak. Actually, the way Hitchcock and Sober set up things does not differ from the intuitive examples in the previous section. Hitchcock and Sober (2004, pp. 16-19) argue that the predictivist advantage comes into effect specifically in cases where we do not have information about

the contents of the model at hand (and do not know the methods that were used in selecting it). In these cases, novel predictive success can provide some guarantees that overfitting has not occurred. Yet, publicity of results is a standard requirement in science (see, for example, Howson 1988). Scientific publications generally include a methods and a results section, where the respective methods and the results of the study are reported. As far as we have the methods and the results explicitly on the table, Hitchcock and Sober do not argue for a predictivist advantage. On the contrary, they hold that after appropriate measures have been taken to avoid overfitting (e.g. AIC), *accommodation* is superior, because it is better to use more rather than less data in fitting our model (ibid. pp. 17-18).

Hitchcock and Sober are, of course, well aware of the modest nature of the predictivism they have presented (ibid., p. 20). They have offered a general template, and sought to demonstrate how a predictivist advantage can arise under certain conditions. To this effect, it will be more useful for us to evaluate the overfitting argument from a broader perspective. For one, even if transparency with results and methods is the ideal, scientists may fall short of this, and there may be differences in the level of reporting detail that are required in different fields. We should seek to understand what the significance of Hitchcock and Sober's argument is given that scientific practice is not perfectly transparent. For another, Douglas and Magnus (2013, pp. 583-584) have argued that overfitting is an even more general problem in cases where we are making inferences from data to phenomena than Hitchcock and Sober recognize. Hitchcock and Sober appeal to the Akaike Information Criterion (AIC) in particular to argue that accommodation is better if we have the model explicitly on the table, but Douglas and Magnus argue that whether the assumptions that underlie AIC hold is often not clear. For this reason, they argue that *temporally* novel predictive success is better than accommodation with AIC. Finally, AIC is not the only statistical method that is used to combat overfitting. Other important methods include the Bayesian Information Criterion (BIC) and cross-validation. The relationship between overfitting, different methods of combatting overfitting, and the prediction versus accommodation issue will require some further clarification.

To ease the evaluation of the pertinent issues, I will consider two separate questions concerning overfitting and the negative predictivist thesis. First, we should seek to understand the relationship between statistical methods of combatting overfitting and the prediction versus accommodation issue. Hitchcock and Sober have argued that accommodation with AIC is preferable to novel prediction. Yet, AIC does not apply universally, and it is certainly not used universally in all scientific fields (see, for example, Yarkoni & Westfall 2017). What do other statistical methods reveal about the predictivist position? Second, do the arguments of Hitchcock & Sober and Douglas & Magnus establish an advantage of novel prediction over accommodation? Hitchcock and Sober argue for an advantage in contexts of missing information about the model and the methods, while Douglas and

Magnus hold that the advantage is there in any case where the situation is not entirely familiar. I suggest that we unify these arguments as 'the argument from unclarity:' accommodation risks overfitting the data in situations where we are not entirely clear about the model, the methods, or the underlying assumptions. Our primary problem in this section is to evaluate whether the argument from unclarity holds in science, and what is its importance for the negative weak predictivist case.

Let us begin with the model selection methods. Steele and Werndl (2016) provide a useful comparison of different model selection methods that are used to evaluate overfitting. We have already encountered the AIC, which measures a model's fit-to-data but punishes it for complexity (see section 2.3). The purpose of the AIC is to select the model that has the highest expected predictive accuracy. In contrast to AIC, the BIC attempts to identify the model with the highest marginal likelihood. The formula for BIC is the following (see Sprenger 2013, p. 107):

$$BIC(M\theta, x) = -2 \log p\left(x \middle| \hat{\theta}(x)\right) + K \log N$$

The first part of the formula contains the log of the likelihood for the maximum likelihood estimator (MLE) for the unknown variable θ, while the latter part is a correction term that corrects for an overestimation of the marginal likelihood in a way that is dependent on both the number of free parameters (K) and the number of data (N). The formula for BIC is thus otherwise similar to the AIC, but its penalty term for the number of parameters also takes into account the number of data. Finally, cross-validation methods are used to estimate a model's long-run predictive accuracy. The standard procedure of cross-validation is to randomly split the dataset into k parts, and then fit a model to k-1 parts. The fitted model is then used to predict the holdout part of the data, and its predictive accuracy is recorded. This procedure is repeated multiple times with different partitions of the dataset, so that each k_i part of the data take turns in being the one predicted based on the rest of the data. An average of predictive accuracy is calculated to estimate the model's expected predictive performance with new data. Minimally, the holdout data that is not used to fit the model contains only one data point (i.e. n-1 cross-validation). In this case, the cross-validation procedure is repeated one by one with each data point.

The alternative model selection methods are based on somewhat different starting points and assumptions, and the choice between them will depend intimately on the context and the aims of the modeler (see Arlot & Celisse 2010). However, they each share a common purpose: to address the problem of overfitting. When we are modeling a dataset, we look to avoid two undesirable outcomes. On the one hand, we do not want to miss the actual information that the data contain. To avoid this, we add parameters into our model to achieve a closer match between our model and the data. One the other hand, we do not want to be too flexible with the idiosyncrasies of our data so that we accidentally capture patterns that are not there. To avoid this,

we must keep parameters at a minimum. The inherent tension between these two goals is why model selection methods are useful and needed. The model selection methods enable us to test more or less flexible models, and thus help us evaluate where the sweet spot between too much or too little flexibility lies.

What do the model selection methods reveal about the predictivist position? First, we should observe that in all of these methods, all of the data are used in building and evaluating the model. In all the methods, the data at hand is taken advantage of in selecting the optimal model. In other words, we do not find support for the standard use-novelty criterion, according to which data that has not been used in constructing the model provides stronger confirmation. Steele and Werndl (2016, 2018) argue, however, that model selection methods reveal nuance in how rules about use-novelty work in science. There is a sense in which cross-validation methods do reserve a role for a certain type of novelty-rule: in cross-validation, data points take turns in being in a novel and accommodated role. Data points are temporarily allocated to an out-of-sample role to gauge the predictive performance of the model. In contrast, AIC and BIC use all of the data at once, but they achieve a similar end-result by penalizing overly flexible models.

What, if anything, can we learn about the predictivism issue based on the more nuanced view of model selection methods? I suggest that the chief lesson here is that questions about novelty-rules break down to even more detailed level than has been previously recognized in the philosophical literature. The overall result with regard to (standard) use-novelty is that statistical model selection methods do not uphold the contingent use-novelty rule: in all model selection methods, all of the data are used both in constructing and evaluating the model. However, in one of the methods – cross-validation – long-term predictive accuracy is estimated via a particular methodological procedure, where data points take turns in being assigned to a transitory, locally novel role. The fundamental purpose of this procedure is to make the most of the data that we have. As far as we are interested in evaluating our model specifically for how well it is likely to perform predictively (versus how well it represents the particular data at hand), this procedure enables us to obtain good comparative estimates without the need to resort to gathering more data. I suggest that we take this as one data point in considering our overall problem of the value of prediction in science. What can be said about model selection methods is that, drilling down to the level of modeling a particular dataset, there is a particular kind of novelty-rule that applies in cross-validation, where it is acceptable to recycle your novel and accommodated data points. This is very different from the kind of novel prediction that concerns the prediction of data points that were not used in the construction of the model (cf. Douglas & Magnus 2013, p. 584), let alone more general scientific inferences outside the statistical context. I will return to this issue again briefly in Chapter 6, where I evaluate more closely what kind of an overall account on the predictivism issue can best capture the various aspects of the role of prediction in science.

Let us then turn to our main problem, the overfitting argument from unclarity. Does the problem of overfitting indicate that we should prefer novel prediction to accommodation in scientific contexts? Hitchcock and Sober (2004) argue that we should if we do not know the contents of the model at hand and the methods that may or may not have been used to combat overfitting. Douglas and Magnus (2013) suggest more broadly that there can still be advantages to novel prediction, particularly temporally novel prediction, even if we know the model and the methods, if we are unsure whether the assumptions underlying our formal methods of combatting overfitting hold (or whether we are applying them properly). Are these arguments successful in justifying an advantage of novel prediction over accommodation for scientific evaluators?

I suggest that, interestingly, and contrary even to what some anti-predictivists have thought (e.g. Harker 2008), these arguments are generally *un*successful. Even with the assumptions of Hitchcock and Sober (2004) or Douglas and Magnus (2013), there are many situations where novel prediction either has no advantage over accommodation, or accommodation is even superior to novel prediction. Novel prediction could under certain limited circumstances provide some additional advantage, but this advantage is not very strong, and it can also be exceeded just by more accommodation. To see where the issue with the predictivist arguments lies, we need to expand the picture on model evaluation. In particular, we need to consider two important perspectives to the overfitting issue that have either so far been neglected in the discussion, or have only been alluded to very briefly (see Douglas & Magnus 2013, p. 583).

The first perspective is to consider also other factors that affect the reliability of statistical inference, namely *sample size* and *effect size*. Model selection methods provide one way to combat overfitting, and novel prediction perhaps another. However, it is well known that another very effective way is just to increase your sample size (see Yarkoni & Westfall 2017, p. 1108). Yarkoni and Westfall (ibid.) explain that with larger and larger samples, it becomes increasingly difficult even for flexible statistical models to capitalize on patterns that only exist in the dataset but not in the general population. This is easy to see intuitively: the larger our dataset becomes, the more representative it will be of the population from which it is drawn (i.e. differences between the sample and the population become smaller and smaller). So, in so far as we seek to avoid overfitting, one of the most effective methods is to just *use more data*. Another important factor is effect size. Statistical inference becomes less reliable the smaller the actual effect is in the population (and more data is then needed to increase reliability). As far as the signal in the data is clear, the overfitting issue can also become negligible (see Yarkoni & Westfall 2017, pp. 1102-1103).[50]

[50] What kind of sample sizes or effect sizes are enough to make overfitting negligible? Yarkoni and Westfall (2017, p. 1103) give an example of sample size 200, with 5 uncorrelated predictors, each

The second important perspective to the overfitting issue is comparing the value of novel prediction to other epistemically relevant factors that kick in automatically whenever a new dataset is used. In cases where a model is tested with completely new data, there are at least two advantages that arise independent of the novel prediction vs. accommodation distinction. The first advantage is simply that of having access to *more data*: when we test if our model fits another dataset, we get more evidence about the model and its performance with the data-generation process in the world. For there to be a *predictivist* advantage in specific, it must be better epistemically that we accommodated one dataset and predicted the other rather than simply accommodated both datasets. The second advantage is the benefit of calibrating (or testing) our model with *an independently gathered* dataset. With just about any dataset that has been gathered by a particular researcher or research group (at a particular time and in a particular location in the world), there is a chance of bias that arises from idiosyncratic aspects of the data collection process. Access to another, independently gathered dataset can potentially address some of these issues, irrespective of the prediction versus accommodation problematic.[51]

If we consider the overfitting issue and the negative predictivist thesis from this broader perspective, where we take into account sample size, effect size, the benefit of having more evidence, and the benefit of having access to an independently gathered dataset, whether there is an important predictivist advantage that pertains to overfitting is no longer clear. At the very least, it does not appear very general or powerful. Consider first Douglas and Magnus's (2013) argument according to which temporally novel prediction is better than accommodation with AIC because whether the assumptions underlying AIC are met is often not clear. First, it should be pointed out that if this issue is considered from the point of view of full array of model selection methods available, the problem Douglas and Magnus raise is already significantly mitigated. Whereas AIC is indeed reliant on some rather idealized assumptions, other model selection methods are much more widely applicable. In particular, according to Yarkoni and Westfall (2017, p. 1111), cross-validation, unlike the AIC, "can be applied to virtually any statistical estimation procedure." Furthermore, they state that "the cross-validated estimate of a model's generalization performance will (on average) typically be

correlated 0.4 with the predicted variable. Assuming a case with a "true" linear relationship, the R^2 of the fitted model is only 0.01 off from the true asymptotic value.

[51] For a practical example of the value of independent datasets, see, for example, Liem et al. (2017). Liem et al. use multimodal imaging data to predict brain-age. They use cross-validation to obtain an in-sample prediction error of around 4 years. A test of the model with another, independently gathered dataset (another country, different scanner, etc.) shows a prediction error twice the size of the original sample. However, a better fit with the other dataset is achieved when the model is trained with all of the in-sample data and some of the data from the other dataset. Liem et al. (ibid.) call for models to be trained (i.e. accommodated) with more independently gathered datasets to "…avoid fitting models to the idiosyncrasies of a given study[.]" In other words, there is a specific advantage to using datasets gathered independently by different research groups.

very close to the true out-of-sample performance" (ibid.). In other words, the limitations of the AIC can often be overcome with the use of other model selection methods, and the obtained estimates of predictive performance are often highly accurate.

Second, even if we are not perfectly confident about the assumptions underlying our model selection method, it is not obvious that novel prediction in particular is what increases our confidence in the model when temporally novel, out-of-sample evidence is obtained. Consider two alternative explanations: a) there is now more data, and b) there is the benefit of using another independently gathered dataset. Overfitting is an issue of too much flexibility with patterns in the data in a particular dataset. With more data, particularly coming from diverse, independently gathered samples, this concern becomes less and less pressing. It then becomes very difficult to see why a model that has been fitted to less data should be considered better. Consider the following example. We have two independently gathered datasets, D_1 and D_2, each containing 200 data points. Researcher A fits and cross-validates a model, M_a, to D_1 and then obtains reasonably good predictive performance with D_2. Researcher B uses both datasets, D_1 and D_2, to fit and cross-validate another model, M_b. We know the contents of both M_a and M_b, which are roughly equal in simplicity, and we know that both researchers used cross-validation. In both cases, the total amount of data is equal: 400 data points. In both cases, we have the benefit of two independent processes of data collection. In both cases, we thus have *two* controls for overfitting: 1) cross-validation, and either 2a) novel prediction or 2b) more data. In such a setting, I submit that there is no compelling reason for why M_a should be preferred to M_b. First, cross-validation already provides a very good estimate of the expected predictive performance of both models (see Yarkoni & Westfall 2017, p. 1111). Second, by using more (independently gathered) data, researcher B, similar to researcher A, has obtained even further protections against overfitting. Arguably, in so far as there is an epistemic difference here, it is to the benefit of M_b, as whatever the real data-generation mechanism is in the world, it is less likely to be distorted by idiosyncrasies in the sample data the more data is used – as Hitchcock and Sober (2004) argue. In other words, as far as we are considering a model that is presented for public evaluation, and know that appropriate methods have been used to avoid overfitting, I suggest that either no predictivist advantage remains, or there is even an accommodationist advantage, as it is better to use more rather than less data in fitting our model.

The issue becomes somewhat more complex if we specifically deny the evaluator access to information about the model and the methods, as in Hitchcock and Sober's version of the argument from unclarity. However, I suggest that the underlying dynamics remain similar. There are multiple ways in which the concern about overfitting can be addressed without using specifically novel prediction even in the type of cases that Hitchcock and Sober consider. One way to address this concern is access to information that a large

enough dataset has been used (relative to the type of data-generation mechanism that is being modeled). As Yarkoni and Westfall (2017, p. 1108) point out, when there is enough data, even an overly flexible model is unlikely to overfit the data (as far as the process we are modeling is not overly complex itself; for that specific case, see section 4.5). A large effect size can give similar protections. Consider, for example, that we have a standard linear regression model, where the relationship between the independent and the dependent variable is strong. Provided there are no concerns about the number of data, overfitting will not be a significant problem in this case. Varying these factors shows that situations can also arise where we have reasons to prefer more accommodation to novel prediction. Take, again, a case with a relatively simple regression model. Consider that we are given the option to select either a) a model that was fitted to a sample of 100 data points and then used to obtain reasonable predictive success with a sample of another 100 data points, b) a model that was fitted to 200 data points, c) a model that was fitted to 300 data points, or d) a model that was fitted to 500 data points. Arguably, d is the best option on the table, as whatever the real data-generation process is in the world, it is much more likely to be adequately represented in a large rather than a small dataset. However, c and even b could be preferable to a, as doubling or tripling a relatively small sample size can result in meaningful improvements in model performance.[52] In other words, even if we do not know the contents of the model or which specific methods were used, more accommodation could be advantageous to novel prediction, as using more data will likely result in the model becoming more accurate.

If we put these observations together, I suggest we are able to contextualize Hitchcock's and Sober's weak predictivism in a way that is revealing about the importance of the negative predictivist argument in this context. Given the points above, we can see that in so far as there is basis for a predictivist argument in this context, this requires that a) we have low information about the model and the methods in question, b) the sample that has been used to fit the statistical model is small enough to raise concerns about overfitting, and c) it is unclear that there is a real effect in the world to begin with (e.g. the effect size is small). In other words, if our epistemic situation is insecure or uncertain to start with, there might be some basis for making a modest predictivist argument. Novel predictive success might

[52] A comparison between an unknown model that has been fitted to 100 data points and then used to predict another 100 data points versus an unknown model that has been fitted to two independent sets of 100 data points is particularly tricky. The former model gains some inductive support about being able to overcome overfitting based on its predictive success with the other dataset. However, we also know that a model that has been fitted to more data is likely to be more accurate with the data-generation mechanism in the world, which benefits the latter model. I suggest that there is no universal solution to which model is preferable. This is a matter that may depend on further background knowledge in the situation. If the existence of the data-generation mechanism is unclear based on background knowledge, the former model may be slightly advantageous, as there is a demonstration of its ability to overcome overfitting. However, if the data-generation mechanism is independently plausible in light of background knowledge, the model that has been fitted to more data could be preferable, as it is more likely to better represent the real data-generation mechanism.

alleviate some concerns about overfitting over and above simply the benefit of having access to more evidence in some of these type of cases.[53] If, however, our epistemic situation is not so insecure – which can be specifically achieved with more accommodation (i.e. by using more data) – the predictivist advantage begins to disappear, and even turns into an accommodationist advantage. Access to more evidence becomes advantageous, as it builds greater and greater confidence in that we are dealing with a model that adequately represents a real data-generation mechanism in the world.

Far from being an important or powerful method to deal with the overfitting problem, novel prediction appears relatively insignificant, all things considered. Looking at the different options for how to deal with overfitting, novel prediction does not appear very important among the available options, and it is certainly not the best option.[54] Arguably, the best option is just to increase sample sizes (i.e. more accommodation). This is something that Hitchcock and Sober also recognize in that they hold that as far as appropriate model selection methods are used, it is better to use more rather than less data in fitting our model. Yarkoni and Westfall (2017) concur. Their suggested remedies to combat overfitting in science are: 1) increase sample sizes, 2) use cross-validation, and 3) constrain models to prior knowledge. (Attempting novel prediction does not make the list at all.) Another good option revealed by our discussion is increasing transparency with regard to the model and the model selection methods: when the model and the methods are reported transparently, evaluators are able to gain much more reliable information about the expected predictive accuracy of the model.

To conclude, we have not seen compelling evidence for that novel predictive success is an important epistemic indicator in science because of the overfitting problem. When other epistemically relevant factors are taken into account (sample size, effect size, the benefit of adding more evidence, and the benefit of using two or more independently gathered datasets), the advantage of novel prediction either becomes very small, or in many cases disappears altogether. A better way to combat overfitting in science appears to be to just use more data, use appropriate model selection methods, and increase the transparency of scientific research.

---

[53] In these uncertain circumstances, it may still be very difficult to establish that novel prediction in particular is better to simply having access to another independently gathered dataset. The epistemic reality may be quite complex in these situations, and I doubt that there is a universal solution to whether novel prediction or accommodation is advantageous in these cases (see also previous footnote).

[54] Here is another way to see the relative impotence of novel prediction in addressing the overfitting problem. In predictivist examples such as Hitchcock and Sober (2004), the predictor is assumed to achieve very good predictive success with new data. However, it is questionable how often scientists would actually achieve such high predictive success, particularly if they originally use a small sample, which is when novel predictions could become more relevant. It may be more likely that the novel predictions eventually fail or turn out much less impressive than what the original sample gives hope for. In other words, to rely on novel prediction to combat overfitting is to rely on an unreliable method to begin with.

4.3 HYPOTHESIS HUNTING, P-HACKING, AND OTHER QRPS

Another prominent negative weak predictivist argument is Mayo's (1996) hypothesis hunting argument. Mayo argues that after-the-fact hypothesis generation decreases the reliability of statistical inference, so pre-trial hypothesis construction should be preferred in the context of statistical experiments. In a typical statistical experiment, a hypothesis is proposed in advance, data is collected to test that hypothesis, and a statistical significance test is performed to evaluate how likely it is that the patterns in the data have arisen by chance rather than represent a genuine effect in the target population. A threshold for an acceptable probability of chance that is used in many scientific fields is $p = 0.05$. This means, roughly, that if we were to draw repeated random samples from the target population, we would expect to see the kind of pattern that we see in the data arise by chance in ~5 % of cases given that the *null hypothesis* is true (i.e. the pattern we are testing for is not there). Clearing this threshold is consider enough to report a result as *statistically significant*, i.e. there is a relatively low probability that the result is due to mere chance.

If the threshold of statistical significant is set relatively high (e.g. 0.05 rather than, say, 0.00003, as in certain experiments in the natural sciences), random patterns do clear the threshold relatively often (with the number of trials conducted). This gives rise to the hypothesis hunting problem. The problem occurs if researchers gather a dataset with a large number of variables, and then simply search through the data for all possible combinations of large, statistically significant effects between variables. Finding some, the researchers suppress all the non-significant results, and report the significant results as if they had been testing for those effects all along. The issue is not difficult to see: by testing more and more variables, it becomes increasingly more likely that *some* relationships between *some* variables are going to show statistically significant effects merely by accident. Thus, by snooping through the data, the researchers end up increasing considerably the probability that they capitalize on chance.

Mayo argues that the problem of hypothesis hunting justifies a preference for novel prediction over accommodation in the context of statistical experiments. When hypotheses are designated prior to statistical experiments, they are tested at the appropriate level of statistical significance, and the hypothesis hunting problem is avoided.

If taken in isolation, Mayo's argument certainly appears compelling: it is undeniable that increasing the number of trials increases the probability that chance patterns are discovered. However, there are crucial pieces of context that are still missing from the discussion. First, hypothesis hunting is not the only methodological problem that arises in the context of statistical tests. Scientists have identified multiple issues, called *Questionable Research Practices* (QRPs), many of which apply particularly to novel

prediction instead of accommodation. To evaluate the strength of the hypothesis hunting argument for the negative predictivist case, we need to expand the picture and consider the problem from the point of view of QRPs in general. Second, Mayo is tacitly presuming that when patterns emerge in the data by chance, a convincing (false) theory that can (often) be developed about those patterns. In other words, she is making assumptions about *scientific background knowledge*. These assumptions warrant more scrutiny. Third, Mayo considers the hypothesis hunting problem by focusing on a single scientist or research group. However, science is a communal practice, where multiple scientists and research groups conduct multiple statistical tests seeking to publish results. We also need to consider what effect *multiple tests within a community* have on the probability that published results are real.

Let us start with the assumptions about background knowledge. Mayo has presented her argument assuming, in effect, that if researchers uncover unhypothesized patterns in the data, it is typically likely that they can use those findings to construct false research hypotheses that will appear plausible in the scientific context at hand. This assumption, however, is by no means always given, and recognizing this enables us to put a certain bound on Mayo's argument. Whether or not Mayo's argument goes through depends on not one but two factors that together regulate the reliability of statistical inference in a given scientific field. The first factor is the acceptable probability of chance that scientists adopt in statistical significance tests. This is what Mayo has focused on, and this is what statistical significance tests measure. The second factor is the probability that any given *plausible looking theory or hypothesis* (that can be formulated based on an unhypothesized pattern in the data) is true rather than false. This, on the other hand, is an issue that intimately concerns scientific background knowledge, and its ability to distinguish between true and false hypotheses.

In order for Mayo's argument to work, certain assumptions must hold about each factor. First, the probability of chance that the scientists accept must be high enough that spurious, unhypothesized patterns can be expected to clear it given the number of trials conducted. In many fields, the designated threshold is 0.05, which certainly meets this assumption. However, this alone is not enough to cause problems for scientific inference. In order for unhypothesized patterns to constitute a problem, scientists must also be unable to recognize whether such unhypothesized patterns are more likely to be real or spurious when they do appear. Any unhypothesized but statistically significant false pattern does not constitute a problem as long as it can easily be ruled out as spurious. Consider, for example, that we have conducted a study on the association between self-efficacy and a number of psychological variables, and the data happens to show a correlation between generalized self-efficacy and indecision. In light of background knowledge about human psychology, this result appears clearly spurious, and it would be hard for us to present it as a plausible hypothesized effect in a psychology journal. Similarly, if the data show an association between self-efficacy and higher age within a

particular age group (e.g. 30-year-olds), the result appears more likely spurious rather than real. In this way, even if unhypothesized patterns do emerge, many of them could be easily dismissible based on background knowledge.

Conversely, any unhypothesized but statistically significant pattern is also not a problem if it is actually *real* rather than spurious (and can be effectively identified as such). The flip side to Mayo's argument is that the search for more patterns in the data also increases the probability that more true effects are found (see, for an example, Hollenbeck & Wright 2017, pp. 6-7). The more scientists look for patterns, the greater the probability that they stumble on more real patterns. The unhypothesized, statistically significant effects that emerge in datasets become a problem to scientific inference *only if* scientists cannot effectively weed out which of these patterns are more likely real and which are false. And, whether or not they can do this depends intimately on the state of scientific background knowledge.[55]

My purpose in introducing this angle to the hypothesis hunting argument is not to argue that Mayo's argument is mistaken. Rather, the purpose is to clarify what this argument depends on, and highlight the possibility that it may not come into effect in all cases or in all fields of science. Mayo's argument depends crucially on an assumption that I suggest we call *the accommodative plasticity of scientific background knowledge*. In order for hypothesis hunting to become a problem, we need a scientific field where the constraints on hypothesis construction are relatively lax. If datasets in some field are likely to contain a relatively large number of spurious patterns that are nonetheless able to masquerade as plausible effects in light of prevailing background knowledge, there are more opportunities for researchers to hunt for and publish those spurious results. In contrast, if the standards for hypothesis construction are stricter, it can be much less likely that scientists are able to find and publish spurious but plausible looking findings.

The assumption of the accommodative plasticity of scientific background knowledge indicates that there could be meaningful variability between scientific fields on whether hypothesis hunting constitutes a problem. Is there anything we can say about which fields are more likely to be implicated? In an interesting discussion, Muthukrishna and Henrich (2019) argued recently that the fields implicated in the recent replication crisis (see below), e.g. psychology, have a problem that corresponds to what could be equated with our notion of accommodative plasticity of background knowledge. They argue that these fields suffer from a lack of generally accepted

[55] In effect, we are now considering the *base rates* of plausible true and false effects in a particular field. With a high base rate of plausible true effects, the probability that a plausible, unhypothesized statistically significant effects is real is quite high. Conversely, with a low base rate of plausible true effects, the opposite is the case. Ulrich and Miller (2020) argue along these lines that the chief reason for the replication crisis in the social sciences is the low base rate of true effects among research hypotheses in multiple fields.

theoretical frameworks that constrain hypothesis construction. Hypotheses are constructed based on personal intuitions and culturally biased assumptions, which leads to a proliferation of disjointed results. As far as Muthukrishna and Henrich are correct, we might consider these fields as more likely to meet the assumption of accommodative plasticity of background knowledge (to a certain degree). The hypothesis hunting argument thus plausibly comes into effect in these fields. In contrast, Muthukrishna and Henrich argue that this is less likely to be the case in fields such as physics, chemistry, and biology, where there are general theoretical frameworks that significantly limit which hypotheses can be considered plausible to begin with. These fields, in turn, may have fewer issues with hypothesis hunting.

For the purposes of what follows, we do not need to take a stance on where the relevant distinction between the fields lies. I propose that we recognize accommodative plasticity of background knowledge as *a constraint* on Mayo's argument, and move to examine more closely the contexts where the hypothesis hunting problem does most plausibly come into effect, i.e. fields where background knowledge is (arguably) accommodatively plastic. This gets us to the second issue against which Mayo's argument needs to be evaluated: Questionable Research Practices (QRPs).

Questionable Research Practices have generated an enormous amount of discussion in the social and medical sciences in recent years (see, for example, Simmons, Nelson & Simonsohn 2011; John, Loewenstein & Prelec 2012; Open Science Collaboration 2015; Head et al. 2015; Bosco et al. 2016; Shaw 2017; Yarkoni & Westfall 2017; Hollenbeck & Wright 2017; Rubin 2017, forthcoming; Vancouver 2018; Murphy & Aguinis 2019). Recent attempts to replicate research findings have shown that many important discoveries turn out spurious under closer scrutiny. This has prompted scientists to search for reasons that have led to such low replication rates in their fields. Among the potential culprits, scientists have pointed to various QRPs and methodological issues, including overfitting (e.g. Yarkoni & Westfall 2017), hypothesis hunting or "question trolling" (e.g. Murphy & Aguinis 2019), p-hacking (e.g. Head et al. 2015), HARKing (e.g. Bosco et al. 2016), suppression of falsified hypotheses (e.g. Rubin 2017), small sample sizes (e.g. Bakker, van Dijk & Wicherts 2012), selective publication (e.g. Bakker et al. ibid.), etc. To be able to further evaluate Mayo's negative predictivist argument, we need to parse through these various issues and isolate the relevance of novel prediction and accommodation, as they relate to QRPs.

The first thing that one encounters in considering QRPs based on the scientific literature is that there is considerable variety in the terminology with regard to QRPs. It is not difficult to find research articles where what are described as the same QRPs are given different monikers and combined in different ways with other QRPs. For example, what Head et al. (2015) identify as p-hacking is called HARKing by Murphy and Aguinis (2019) while Yarkoni and Westfall (2017) name it procedural overfitting. To clarify the discussion and make possible a comparison of *novel prediction and accommodation* in

particular, I propose that we classify the various QRPs identified by the scientists as follows, based on the use-novelty vs. accommodation distinction:

*Hypothesis hunting (*or, HARKing*)*. In terms of the prediction versus accommodation distinction, what scientists call HARKing comes closest to what has been debated in the philosophy of science. HARKing is short for "hypothesizing after results are known" (see Kerr 1998), i.e. accommodation. In what follows, I will refer to HARKing in the strict sense where it is equivalent to accommodation, i.e. *constructing* hypotheses based on known results (called "CHARKing" by Rubin [2017]). The epistemic issue with HARKing, as far as there is one, is what Mayo (1996) calls hypothesis hunting and others have called "question trolling" (e.g. Murphy & Aguinis 2019). The concern is that the more researchers search for effects in the data, the greater the probability that they end up capitalizing on chance and either promote false hypotheses or inflate real effect sizes. HARKing as such is not necessarily always a problem, even in fields where background knowledge is accommodatively plastic (see Mayo 1996, pp. 294-318; Hollenbeck & Wright 2017; Rubin 2017). If, for example, researchers gather a large dataset and happen to stumble on one strong, independently plausible effect, using that effect to construct a new hypothesis may not constitute much of a problem. The issue starts to emerge when researchers *systematically* look for unhypothesized effects in the dataset.[56]

P-hacking. In contrast to hypothesis hunting, scientists have raised another set of issues called p-hacking. P-hacking is essentially the reverse of hypothesis hunting. In p-hacking, the researcher starts with a novel prediction (i.e. a research hypothesis from which a novel consequence is derived), and then performs various manipulations or adjustments on the data or statistical tests to ensure that a statistically significant result that appears to support the hypothesis is achieved. This corresponds to concerns that anti-predictivists such as Dellsén (forthcoming) and Harker (2008, p. 440) have raised about novel prediction (see also Brush 1989, p. 1127; Lipton 2004, pp. 176-177). The underlying issue is the same as that in hypothesis hunting: if the researchers make enough adjustments on the data and the tests to bring the data into agreement with theory, the probability that either false or inflated results end up promoted increases. Types of p-hacking identified in the scientific literature include trying out several statistical tests to find one that shows a significant result, analyzing data midway through data collection to decide whether to continue collecting more data, including or excluding

[56] In a useful discussion, Rubin (2017) further distinguishes between 'CHARKing,' 'RHARKing' and 'SHARKing.' CHARKing refers to what is understood here as hypothesis hunting: using the data at hand to construct hypotheses after statistical trials. In contrast, RHARKing means searching the scientific literature for existing hypotheses that are supported by the data at hand. As Rubin (2017) correctly points out, the data in this case provides a use-novel test to the RHARKed hypothesis. RHARKed hypotheses are thus appropriately excluded from what are considered here as the potential issues of HARKing as accommodation. What Rubin calls SHARKing is considered in this study under the issue of 'evidence suppression.'

outliers, including or excluding covariates, and stopping data collection if statistical significance is reached (see Head et al. 2015).

Evidence suppression. In addition to p-hacking, certain issues of evidence suppression have also been raised in the scientific literature. These could also be seen as implicating novel prediction in a problematic way. One issue is dropping novelly predictive hypotheses from research papers if they fail to show statistically significant results. Curiously, this is often called a form of HARKing by the scientists (see, for example, Bosco et al. 2016; Rubin 2017). However, from the philosophical point of view where we distinguish between (use-)novel prediction and accommodation, this is clearly not an issue where evidence is used to construct a hypothesis. Rather, we are talking about the failure to report the falsification of a novel prediction. The failure to report such disconfirmations constitutes a problem particularly at the *community level*, which is a point that we shall consider at the end of this section.[57]

Another potential problem that arises from evidence suppression is the conduction of multiple studies to try to confirm a novelly predicted effect (see Ulrich & Miller 2020, p. 5). In these cases, the researcher postulates that a certain effect exists in the population, conducts multiple studies going through multiple disconfirmations, and then finally reports the result of a study that does appear to confirm the hypothesized effect. In other words, the researcher actively suppresses evidence against his novel hypothesis. This could be called p-hacking with a vengeance. Unfortunately, this issue is not typically distinguished from other QRPs in scientific research articles that estimate the prevalence of p-hacking, so I am unable to estimate the extent of this problem in science. However, given the prominence given to other types of p-hacking in the scientific literature, this practice could be relatively rare.[58]

Fraud. Finally, there is also the problem of outright fraud. In principle, as long as we are talking about a single study conducted by an isolated researcher or research group, it is possible that the entire study is based on fraud – most prominently, fabrication of data. Some anti-predictivists such as Dellsén (forthcoming) identify fraud in particular as a

[57] There is one sense in which this practice could negatively affect the confidence we have in novelly predictive hypotheses also at the level of individual studies. Rubin (2017, pp. 316-317) raises the possibility that if a researcher starts with, say, 20 novel predictions on a particular topic, ends up confirming only one, and then suppresses all the others, we appear to have the same problem that arises with hypothesis hunting. One could count this as an additional issue with novel prediction on top of p-hacking. However as far as I am aware, no scientist has raised this as a prevalent issue in scientific practice. The effect is most likely considerably stronger at the community level, a point which I will concentrate on in this study.

[58] The self-admission rate for "selectively reporting studies that worked" is quite high in fields such as psychology (see John et al. 2012; Motyl et al. 2017). However, this is done for multiple reasons, including simply due to errors in the experiment (e.g. Motyl et al. 2017, p. 41). Actively suppressing multiple disconfirmations of one's hypothesis indicates a more pernicious activity knowingly to hide problematic evidence. In a meta-analysis of QRPs by Fanelli (2009), the self-admission rate for publishing results which the researchers knew to be untrue was around 1-2 %. This could provide some indications that the purposeful suppression of multiple disconfirmations of one's published novel results is relatively rare.

problem with novel prediction. This, of course, is accurate in the sense that you first need a hypothesis before you can begin to fabricate data to "support" that hypothesis. However, based on the scientific literature, and contrary to Dellsén's assumptions, this problem appears relatively minor in comparison to the other QRPs: outright fabrication is apparently very rare in science. According to Ulrich and Miller (2020; see also Fanelli 2009; Stroebe et al. 2012; Gross 2016), the incidence rate is most likely smaller than 2 %. Even if fraud does occur, it may happen so rarely that it is not enough to raise significant concerns about novelly predicted effects overall. For this reason, I will also pay less attention to fraud in particular as an issue in considering our main problem.

In addition to these four categories of QRPs, there are other methodological issues that likely affect replication rates; namely, overfitting and small sample sizes. I will not consider these further in the context of the hypothesis hunting argument, as these are issues that arise equally regardless of the point of time when we theorize about statistical results. Whether we are making novel predictions or theorizing based on results, it remains the case that it is better to use more rather than less data and take appropriate measures to combat overfitting. (For discussion on the prediction vs. accommodation issue on the level of inference from data to phenomena, see the previous section.)

With these considerations, we are now able to effectively distill our main problem to a comparison between hypothesis hunting and p-hacking, with perhaps some contributions from fraud and evidence suppression on the latter side. The former is a potential issue that implicates accommodation unfavorably. The latter covers a family of issues that concern novel prediction.[59] In both cases, the alleged problem is the same: the probability that spurious or inflated results are promoted increases. There is evidence that both types of practices are used in scientific practice. Based on multiple studies conducted among psychologists (and one among management researchers), Rubin (2017, p. 309) tallies a self-admission rate of 43 % where researchers admit to having HARKed "at least once." Interpreting this figure is somewhat complicated, as the definition of HARKing varies in the scientific literature. Sometimes it includes the suppression of failed novel predictions (e.g. Bosco et al. 2016), and in some surveys the question is phrased vaguely (e.g. researchers are asked whether they "report that unexpected findings were expected") (e.g. Motyl et al. 2017). On the other

[59] Lipton (2004, pp. 176-177) briefly considers this family of problems with novel prediction, calling it 'observational fudging.' He admits that observational fudging is an issue, but alleges that accommodation is still worse than prediction because in the case of accommodation *both* theoretical and observational fudging are available to the researcher. Lipton is undoubtedly correct that in case of accommodation it is technically possible for a researcher to further engage in observational fudging, e.g. p-hacking. However, given that this constitutes a less attractive subset of hypothesis hunting in general (in effect, we need a researcher who searches through the data and finds a *non-significant* relationship between variables, and then further p-hacks the result so that it turns into a significant one), I think it is safe to assume that p-hacking is *more* of a problem with predesignated hypotheses.

hand, given that HARKing may be considered undesirable, Rubin argues that the self-admission rates may also be deflated. Various types of p-hacking show similar numbers. For example, in a survey of social and personality psychologists, Motyl et al. (2017, p. 39) find self-admission rates where researchers admit to having "at least once" selectively reported studies that worked (84 %), decided to collect more data after looking (66 %), excluded measures (78 %), dropped data after looking at impact (58 %), rounded down p-values (33 %), and stopped data collection early (18 %). Comparable results were found by John et al. (2012). Similarly as in the case of HARKing, John et al. estimate that the real prevalence rate of different forms of p-hacking may be higher than what the researchers report about themselves.

The introduction of the full picture on QRPs now enables a simple argument against Mayo's negative predictivism. Unlike Mayo assumes, there are issues with *both* novel prediction and accommodation. If we learn that a hypothesis was proposed prior to data collection, we have reasons to suspect p-hacking (and perhaps evidence suppression or fraud). If we learn that a hypothesis was proposed based on the data, we have reasons to suspect hypothesis hunting. In both cases, the hypothesis at hand warrants less confidence than what initially appears based on the theory and the reported evidence as such. In other words, both novel prediction and accommodation have issues so, for evaluators, there is no particular reason to prefer one or the other.

This, of course, is still *too* simple. In order to draw an exact conclusion about whether one or the other, novel prediction or accommodation, is preferable, we need more precise information about the relative prevalence, nature, and effects of hypothesis hunting and p-hacking in different fields. Unfortunately, such evidence is lacking at this moment. Most likely, it will also be rather difficult to gather, as there are many interconnected factors that affect the reliability of statistical inference (sample size, effect size, etc.) (cf. Murphy & Aguinis 2019). However, there are at least two arguments that we can make with the evidence that we have. First, we are able to establish that whether novel prediction or accommodation is preferable is a purely contingent issue that depends on multiple contingent factors related to the prevalence of hypothesis hunting, p-hacking, and other QRPs, the effects of the different QRPs, typical effect sizes in the scientific field, typical sample sizes, composition of the research field, etc. In other words, in the absence of further evidence from scientific practice, Mayo's argument does not establish that novel prediction in particular should be preferred by evaluators. In many situations, it is actually possible that novel prediction could be worse than accommodation. For example, if we have a scientific field where researchers are reluctant to HARK, engaging in it only if they can rely on large samples and effect sizes, but they routinely engage in p-hacking, evaluators have a reason to trust accommodation more than novel prediction. In the absence of further information about what specific kind of practices are associated with novel

prediction and accommodation in a particular field, there is no particular reason to prefer *either* novel prediction or accommodation.

Second, in terms of the evidence that we do have from actual scientific practice, there are some indications that neither p-hacking nor hypothesis hunting is likely to lead to very significant detrimental effects, and neither is likely to cause much more issues than the other. In a recent simulation study,[60] Murphy and Aguinis (2019) investigated the comparative effect of what correspond to what we have called hypothesis hunting (what they call "question trolling") and a particular form of p-hacking, searching through the data with alternative measures and samples to find the strongest support for a hypothesized result ("cherry-picking"). They perform multiple simulations, varying the sample size (n = 100-280), the pool of results from which to choose (k = 2-10), the prevalence of cherry-picking and hypothesis hunting (20% - 80% of researchers), and the heterogeneity of the pool of effects from which hunters can draw (ranging all the way up to the entire literature in the field), and estimate how much either practice would inflate a real effect size of 0.20. As one would expect, effect sizes are more inflated the smaller the sample size, the greater the number of results to choose from, and the higher the prevalence of the QRPs. However, as long as the prevalence of cherry-picking and hypothesis hunting remain below 50 %, the differences are not very significant. For example, with an equal prevalence of 40 %, in the worst case scenario of sample size 100 and pool size 10, cherry-picking results in an inflated effect size of 0.258 while hypothesis hunting results in 0.264 as far as the hunters are not looking from a highly heterogeneous pool of effects.

Murphy and Aguinis (2019) warn that hypothesis hunting does have the potential of becoming more detrimental than cherry-picking. At worst, with sample size 100, pool size 10, prevalence rate of 80 %, and where the pool of effects from which to choose includes the entire field, hypothesis hunting could double the real effect size (0.417). In contrast, the worst we could do with cherry-picking at 80 % prevalence rate is a 1.5 times increase (0.316). However, this does not establish a predictivist advantage in the current context. First, Murphy and Aguinis do not include all the potential problems associated with novel prediction. For example, in another simulation study, Simmons et al. (2011) show that another type of p-hacking, stopping data collection if statistical significance is reached, can raise the rate of false positives up to 20 %. If all forms of p-hacking are simultaneously taken into account, the calculation may reveal more negative effects to novel prediction. Second, Murphy and Aguinis's simulation presumes that researchers systematically engage in these QRPs. With this assumption, hypothesis hunting starts to become worse at particularly high prevalence rates (e.g. 60 % or higher). However, it is questionable if actual levels are

[60] Ulrich and Miller (2020) provide another simulation study with comparable results about the relative insignificance of QRPs in causing low replication rates. I will focus on Murphy and Aguinis (2019) here, as their article specifically compares the relative effects of particular QRPs associated with novel prediction and accommodation.

anywhere close to this. As discussed before, Rubin (2017, p. 309) reports that 43 % of researchers in psychology admit to having HARKed *at least once*. In the survey by Motyl et al. (2017, p. 40) participants report that they *rarely* or *never* engage in either hypothesis hunting or p-hacking. As far as these reports are even close to reflective of scientific practice, Murphy and Aguinis's simulations do not reveal much differences between cherry-picking and hypothesis hunting. In fact, the detrimental effects will be altogether small in either case.[61] Finally, Murphy and Aguinis also neglect the influence of background knowledge, and the precise type of HARKing that occurs in different cases. In their simulation, they assume in effect that any statistically significant patterns that emerge in the data could be presented as plausible hypotheses in the scientific literature. However, even with accommodative plasticity of background knowledge, it is unlikely that all such patterns are plausible. Furthermore, not all cases of after-the-fact hypothesis construction are problematic, as discussed above. For example, using an unexpected but strong effect from a large dataset to construct a new hypothesis can be perfectly acceptable scientific practice (see, for example, Hollenbeck & Wright 2017). This further constrains the detrimental effects that hypothesis hunting is likely to have in actual scientific contexts.[62]

In summary, considering the overall picture on QRPs, we have not seen evidence that reliance by evaluators on either novel prediction or accommodation is likely to lead to significantly worse epistemic outcomes in current scientific practice. If reports of the prevalence of QRPs are even roughly accurate, we are effectively talking about small negative effects in either case in many fields. After all of the QRPs have been taken into account, there does not appear to be a significant epistemic difference between novel prediction and accommodation due to the hypothesis hunting problem. However, before we deliver the final verdict on Mayo's hypothesis hunting argument, there is one more perspective that we need to consider to this argument. This is *the community level* view on hypothesis selection and publication. There is an interesting sense in which taking this perspective into account may show that novel prediction actually leads to *worse* epistemic consequences than accommodation in current scientific practice.

Mayo has argued that accommodation constitutes a problem for scientific inference because hypothesis hunting by individual researchers

[61] For example, if we assume a systematic prevalence rate of 20 % for both cherry-picking and hypothesis hunting, a sample size of 100, and a pool size of 10, and an actual effect size of 0.2, cherry-picking leads to an inflated result of 0.229 while in the case of hypothesis hunting the result is 0.232, with a further 0.02 increase if we allow for the hunters to sample haphazardly the entire literature in the field.

[62] Another study that estimated the problematic consequences of HARKing is Bosco et al. (2016). They find that hypothesized effect sizes between certain variables in organizational research are around 0.06-0.08 greater than nonhypothesized effect sizes, and attribute this inflation to HARKing. However, Bosco et al. (2016) do not include or study alternative explanations that concern the problems of novel prediction: the greater hypothesized effect sizes could also be due to various types of p-hacking (or selective publication). Accordingly, I suggest that this study fails to show evidence about the relevant distinction between novel prediction and accommodation.

increases the probability that spurious findings are promoted. However, there is also another way in which essentially the same effect can emerge. A version of the hypothesis hunting argument can also be formulated on *the community level*, in a way that affects particularly novel prediction. Objectionable hypothesis hunting can occur on the community level if three conditions are met: 1) scientists actively generate and test novel hypotheses, 2) failed predictions are suppressed, either by the scientists themselves or by publishers who have a preference for surprising, novel results, and 3) true hypotheses are relatively uncommon (i.e. there is a relatively low base rate of real effects in the field). In this kind of a setting, a similar concern as that of Mayo arises: the more novel predictions are tested by the community, the greater the probability that we get hypotheses that pass by chance. If a large portion of hypotheses are false to being with, and this fact remains hidden as falsifications are suppressed, there is a significant chance that a significant proportion of published novelly successful hypotheses are false (cf. Ulrich & Miller 2020).

This point of view has featured prominently in the recent scientific discussion on the replication crisis. In multiple fields implicated by this crisis, scientists have raised the concern that prominent journals have a preference for novel, surprising, statistically significant effects, while null results and replications are given much less space (e.g. Bakker et al. 2012; Antonakis 2017; Woznyj et al. 2018). Based on the low replication rates, scientists have also estimated that the base rates of true effects in these fields are rather low. For example, Wilson and Wixted (2018) estimate that the base rate of real effects is 0.20 in cognitive psychology and 0.09 in social psychology (see also Ulrich & Miller 2020). This generates precisely the conditions that we have identified for a community level hypothesis hunting argument: rather than having individual researchers hunt for significant effects in a particular dataset, we have entire fields of science where scientists hunt for novel, significant results. Failures to find such results are suppressed, and cases where they do appear are promoted. This, as the scientists have observed, can result in significant bias in the literature, as many of the selectively published studies are based on mere chance results.

Arguably, taking into account the selective publication effect may show that reliance on novel predictions is worse overall from the epistemic point of view than reliance on accommodations in the current context. If we have a large community of researchers who hunt for surprising novel effects which, however, are relatively rarely real, novel prediction can become a highly undependable indicator of truth in the field in question. There is evidence that something like this may have happened in the replication crisis. A recent Open Science Collaboration (2015) study that estimated the replicability of 100 psychological studies published in important journals found evidence that *surprising original effects* were among those *least likely* to replicate. In other words, if we used surprising novel success as our standard for what counts as a compelling scientific result, we would very often be spectacularly wrong with

our beliefs. As long as novel prediction has such low reliability, it is possible, even though not guaranteed, that accommodation could be better overall. For example, if scientists are relatively reluctant to HARK, doing it only or mostly when there is a good reason for it, HARKed results could be more trustworthy than novelly predicted results, and thus accommodation advantageous to novel prediction.[63] (The precise verdict here, of course, will depend on further details about the kind of accommodations that individual scientists perform, and remains a matter of speculation at this point.)

To conclude, if we put together the full picture, where we take into account scientific background knowledge, all the QRPs, and the selective publication effect, there is not much evidence for Mayo's argument that novel prediction should be preferred in science due to the hypothesis hunting problem. After all the QRPs associated with both novel prediction and accommodation are taken into account, there is no clear reason to prefer one to the other. As far as differences arise, these are highly contingent on further assumptions about the prevalence of various QRPs, the effects of the QRPs, typical sample sizes and effect sizes, etc. After the community level selection effect is taken into account, it is also plausible that, in the current context, novel predictions are worse overall.

A particularly striking result of our discussion is that similarly as with the problem of overfitting, the novel prediction vs. accommodation framework does not to present solutions to the problems that have been raised here. Promoting either more novel prediction or more HARKing does not resolve the problems that we have uncovered. Rather, the solution to more reliable statistical inference in science seems to be either better methodological practices, or just *more evidence*, e.g. studies with larger datasets or the conduction of multiple studies to investigate the replicability of research findings. Results that are based on a single study are not necessarily very compelling whether the results were novelly predicted or HARKed based on the data. Confidence in results increases if they are replicated in independent studies, as the independent studies are able to control for the potential downsides of both novel prediction and accommodation. This provides some preliminary indications that something may be wrong with the comparison between novel prediction and accommodation in the philosophy of science: there appears to be something else that is more valuable in science than either novel prediction or accommodation. We will return to evaluate this finding further in Chapter 6.

[63] The survey responses in Motyl et al. (2017) of social and personality psychologists indicate that researchers are relatively reluctant to HARK (i.e. report unexpected findings as expected) in the current environment. The practice is seen as "slightly unacceptable," similarly to most forms of p-hacking, and the survey responses indicate that it is done relatively rarely. The survey also includes reasons that researchers give for HARKing when they do engage in it. These include: "a more careful literature review would have yielded the hypothesis," "it was followed up with additional studies and it replicated," editors'/reviewers' strong suggestion," and "when presented as an alternative and theoretically grounded hypothesis" (ibid. p. 41).

4.4 FUDGING

So far, our investigation of the overfitting argument and the hypothesis hunting argument has failed to show an important negative predictivist advantage in scientific practice. Yet, one negative predictivist argument still remains: Lipton's (2004) fudging argument. Lipton argues that accommodation may cause scientists to fudge their theory or auxiliaries in the effort to produce a fit with the evidence. Fudging results in a less simple theory, which provides an inferior explanation of the evidence. Information about the possibility of fudging should therefore lead evaluators to lower their confidence in the theory. Can Lipton help rescue the negative predictivist case and demonstrate a predictivist effect in science after all?

Interestingly, something very much like the fudging argument has been discussed recently in the scientific literature on the replication crisis (see Shaw 2017; Rubin 2017, pp. 313-314; Hollenbeck & Wright 2017, pp. 14-15). Writing from the editorial perspective in management research, Shaw (2017) argues that HARKing often results in manuscripts that contain contorted theories, conceptual sloppiness, and mismatches between the theory and the operationalized measures. Theorists who HARK end up invoking multiple theoretical frameworks in an incoherent way, they define their concepts in a loose way, and they struggle to fit the measures used in the study to the newly chosen HARKed theoretical framework (which implicates other measures and mechanisms more directly). All of this, I suggest, sounds very much like Lipton's fudging argument: HARKed hypotheses may become more convoluted, which decreases the probability that they are successful. In another discussion that raises similar points, Hollenbeck and Wright (2017, p. 14) write that HARKing results in manuscripts where the "[i]ntroduction sections are filled with tortured text that does not make any sense to the well-informed reader." According to them, the reader's reaction to such text is often "confused" and "incredulous."

The scientists' observations provide credibility to Lipton's argument that fudging can be a methodological problem with accommodation in scientific practice. However, this does not yet settle the question of Lipton's weak, negative predictivist argument. In order for Lipton's weak predictivist argument to hold, it is not enough to show that accommodation can lead to the practice of fudging. What matters is that after the theory and the evidence are on the table, does information about prediction and accommodation continue to be relevant with regard to fudging? In other words, do the consequences of fudging – e.g. convoluted theories and constructs – speak for themselves, or do scientists need to resort to novel prediction to avoid the negative epistemic consequences of fudging?

There are, I argue, three issues that point to a negative conclusion against Lipton's argument. At the very least, I suggest that Lipton is unsuccessful in making a compelling case on the behalf of negative predictivism. First, there is the question of actual evidence about the negative

epistemic consequences of fudging. Lipton has not given a single example where fudging occurred but scientists were effectively fooled by it. He has also failed to present any evidence that surreptitious fudging is anything close to a pervasive and general problem in scientific practice. In contrast, the writings of the scientists may provide at least *some* evidence against Lipton's claim. Shaw's (2017, pp. 819-820) take on the dangers of HARKing is framed as a warning to authors. According to him, HARKing is a "rejection-creating" practice that leads to unpublishable manuscripts. He argues that HARKing leaves "telltale signs" that lead to negative reactions from reviewers. Similarly, according to Rubin (2017, p. 314), "the signs [of low quality accommodation] are quite obvious." He explicitly denies the type of predictivist argument that Lipton makes (although not referencing it in specific): "[p]eer reviewers and end-users are able to identify these problems and reduce their confidence in the reported research regardless of whether or not they suspect that researchers have engaged in HARKing" (ibid.). In other words, scientists have expressed confidence about being able to detect the negative consequences of what could be seen as fudging directly.

Second, in contrast to overfitting, small sample sizes, p-hacking, hypothesis hunting, selective publication, and the base rate of true effects in the fields implicated by the replication crisis, the publication of fudged or convoluted hypotheses has not received similar attention in the scientific literature. This does not directly show that fudging is not an actual epistemic issue for the scientists. However, it does provide some further indications that scientists themselves have not regarded fudging as a major problem – despite several years of active discussion about a methodological crisis in multiple fields. If we consider the whole menu of potential contributors to low replication rates (see section 4.3), it is also unclear how much of the low replication rates is left to explain by factors that have not already been covered here. In some analyses, it is argued that much of the low replication rates are already explained by just a subset of these factors (see, for example, Ioannidis 2005; Ulrich & Miller 2020). This, too, provides some further evidence that fudging may not be highly relevant from the epistemic point of view.

Finally, there is also one more principled problem with Lipton's argument. The very nature of the argument implies that whenever fudging occurs, it is *less likely* to result in a theory that has the potential to substantially mislead scientific evaluators. This is because fudged theories are less simple, as they have been adjusted to particular results. Even if such hypotheses end up published, they are restricted to rather isolated contexts. Rubin (2017, p. 313) gives the following example of the kind of hypothesis that we might consider as a result of fudging: "prejudice only increases self-esteem among black women who are aged 50 years or more." Fudging could very well result in the generation of such a hypothesis. However, by its very nature the impact of the hypothesis will be limited. In contrast, the potential for harm will be much greater if we start with simpler, novelly predictive hypothesis that are often false, which is the issue that we identified with novel prediction in the

previous section. With a selection bias that emphasizes impressive (but often spurious) novel findings, the pool of impressive novel findings could lead to much more substantive and comprehensive false beliefs within a scientific field. In other words, even if fudged hypotheses were for some reason able to elude scientific evaluators in some cases, their potential for epistemic harm (e.g. false beliefs) is limited compared to selective publication of impressive but spurious novel findings.

In summary, Lipton's fudging argument for negative predictivism a) lacks evidence from scientific practice, while b) the evidence that we do have points to the opposite conclusion against Lipton's predictivist position. Furthermore, c) Lipton's argument by its very nature limits the problem of bad accommodation, which in the light of previous findings about the problems of novel prediction indicates the potential for epistemic harm may be greater with novel prediction than with fudging.

4.5 EPISTEMIC OPACITY

We began this chapter by discussing intuitive examples that appeared to show a predictivist advantage. This advantage was connected to lack of information about the theories or the independent evidence at hand. We argued that this is not "typically" the situation in science, where the theory and the evidence are reported publicly. However, there is one type of case that has been discussed in the literature where this is not feasible: in certain cases, the theory or the model under evaluation is so complex that it is *impossible* to gain access to its epistemically relevant features. We should finally consider this type of case and its relevance to negative weak predictivism.

The issue of impenetrable complexity and predictivism has been discussed recently in relation to global climate models. In this instance, predictivism has emerged in the scientific literature itself. Given the complexity of their models, climate scientists have endorsed a preference for use-novel data in evaluating climate model 'skill' (i.e. whether climate models are adequate for various predictive tasks) (see, for example, Flato et al. 2013, p. 750; Frisch 2015, pp. 171-172; Baumberger, Knutti & Hadorn 2017). This preference has been defended by philosophers of science (see Frisch 2015; Winsberg 2018, pp. 163-173).

Let us examine the arguments. Frisch (2015) and Winsberg (2018, pp. 163-173) defend the predictivism in climate science by appealing to climate model complexity. Basically, they argue that the predictivist advantage arises in consequence of two factors: *model tuning* and *epistemic opaqueness*. Model tuning refers to the practice of climate model calibration that is done on existing climate datasets. Climate models represent many physical processes that can be modeled based on well-confirmed physical theories, but there are also important processes that are either too small, too complex, or too poorly understood to represent explicitly in climate models. These processes must

therefore be represented in the models in a more inexact way by adding free parameters into the models (cf. Worrall 2006). However, which parameters should be chosen and which specific values those parameters should take is only loosely constrained by background knowledge. So, to find parameter combinations and values that improve model skill, climate scientists calibrate parameter values to existing climate datasets (i.e. they choose parameter values that result in a better fit between existing data and the model's predictions) and experiment with alternative parameterization schemes.

If a tuned model achieves a fit with existing data, it gains an advantage over models that do not fit the data. Thus, the ability to accommodate existing data confirms a model to some degree (see Winsberg 2018 p. 163; Baumberger et al. 2017, p. 8). However, when it comes to the model's predictive skill with data that have not been used to tune it (e.g. future climate data), there is an important limitation to what can be inferred about model adequacy based on accommodative success alone. This limitation arises from the combined effect of model tuning and what Frisch, adopting a term from Humphreys (2009), calls epistemic opaqueness: it is not known or fully knowable how the chosen parameterization scheme contributes to the model's performance. When a model is tuned to reproduce existing data, it is not fully transparent what it is in the complex interactions between model structure and model parameters that result in the model's success with the accommodated data. The success could be due to the fact that the model adequately represents some relevant processes and mechanisms in the climate. Or, the success could also be a fluke of circumstance that arises, for instance, from error compensation by other errors. Because of epistemic opaqueness, it is often impossible to tell what is actually going on.

This is where the predictivist argument kicks in. Frisch and Winsberg argue that if a model succeeds in predictive tasks with data that have not been used to tune it, this provides a reason to think that the model's success might be a little bit more likely due to an adequate representation of relevant climate processes rather than just a fluke of circumstance. That is, predictive success indicates that the model might be slightly more likely to represent some relevant features of the climate in a way that is adequate for further predictions in some pertinent range of similar circumstances. The increase in epistemic support may not be very significant, because it is still not fundamentally understood what is going on in the model. Perhaps, predictive success provides some weak inductive support that the model could be successful in further predictive tasks, as Frisch (2015, p. 185) suggests. Or, perhaps the support is comparative in that it allows for climate scientists to pick models that are at least somewhat better at predictive tasks than alternatives, as Winsberg (2018, pp. 172-173) argues. However, arguably, predictive success is still a better indicator of model skill than accommodative success, because in cases where epistemic opaqueness holds, accommodative success alone does not reveal much about how robust or accurate the model's predictive performance is going to be.

In my view, the arguments by Frisch and Winsberg are convincing. If it is truly impossible to evaluate what makes a particular model successful with certain data, novel data is better in that it can provide at least some test for the model's predictive capabilities. At the very least, we are able to gain some more inductive support for the model, as Frisch suggests. The argument here is, in fact, analogous to the stock market example at the start of the chapter: prevented access to the relevant features of theory or the model at hand, the only way for us to evaluate predictive performance is to observe and measure it. Frisch and Winsberg are thus successful in showing that there are contexts in science where a predictivist advantage arises similarly as in the intuitive examples.

How significant is this to the predictivist case? It does establish something that even anti-predictivists have agreed on, i.e. there are cases in science where a predictivist advantage exists (e.g. Harker 2008). However, in terms of the importance of this advantage, there are significant limitations. First, epistemic opacity is by no means the prevailing condition in science. Any advantage that arises in contexts of epistemic opacity does not translate to an advantage in other contexts, as we have seen in the previous sections. Second, even though novel prediction has an advantage in cases of epistemic opacity, its epistemic force remains *very* modest. Even if an epistemically opaque model succeeds in some predictive task with new data, *we still have no idea why the model is successful*. Predictive success might be lost at any point and we would have no way of anticipating that. This is a feature that climate scientists themselves emphasize about global climate models. Climate scientists stress that climate model projections contain a high degree of uncertainty, and use-novel success does not do much to alleviate that uncertainty (see, for example, Stainforth et al. 2007; Baumberger et al. 2017). Baumberger, Knutti and Hadorn (2017, p. 9) explain that there is a very fundamental issue with the predictive accuracy that we can expect from climate models, one that goes beyond even basic inductive skepticism. This issue arises from model calibration to past and present boundary conditions (e.g. levels of CO_2-emissions). Given that climate models have been calibrated to particular conditions of the climate, but it is not known how the relevant parameterizations behave in other boundary conditions (e.g. higher levels of CO_2-emissions), even models that achieve use-novel success in current conditions can be biased. Baumberger et al. continue to argue that confidence in long-term climate model projections can increase based on their fit-to-data, their fit to other models (i.e. multiple models point to a particular direction), and background knowledge about well-understood physical processes. They also call for mitigating the problem of epistemic opaqueness by performing more detailed experiments that can attribute predictive success to different model components. Novel prediction, again, does not make the list as a promising candidate for boosting confidence in the long-term predictive accuracy of climate models (cf. section 4.2).

To summarize, cases of epistemic opaqueness represent situations where science is the most uncertain. The advantage that novel prediction enjoys over accommodation in these conditions is weak, as stronger forms of epistemic support are still being sought elsewhere (e.g. better experiments, building background knowledge about relevant processes).

4.6 CONCLUSION

Let us take stock. Our investigation into potential negative epistemic consequences of accommodation has failed to show a general or important reason to prefer novel prediction to accommodation in science. After the respective problems of both novel prediction and accommodation are taken into account and other relevant epistemic factors are included, epistemic differences between prediction and accommodation either disappear or are revealed highly contingent on the circumstances. There are indications that reliance on novel prediction may even have more detrimental effects to the reliability of scientific inference than does accommodation in the current context.

We found one exception of a type of situation where novel prediction could provide some advantage over accommodation. This happens in cases where we are either prevented from evaluating theories and methods directly (perhaps in some cases where there are concerns about overfitting) or where this is not possible to begin with (in cases of epistemic opaqueness). However, the value of novel prediction in these cases is limited. Novel prediction provides some inductive support for that we may be on the right track. But, in these cases, novel prediction is not so much a solution to the underlying problem of uncertainty, but a last resort that we may turn to to gain some additional support for hypotheses and models whose epistemic status is otherwise unclear.

The findings of this chapter point, in general, *away* from the novel prediction versus accommodation distinction. We have found that perhaps the most effective way to address concerns about any form of bad accommodation as well as the problems of novel prediction itself is simply *to collect and use more evidence*. This could mean more data in original studies, or replication attempts for original studies. Another important factor is *transparency*. The more transparent the scientific process, the less relevant considerations about hypothesis origin become, and the focus turns to the theory and the evidence as such. The question then becomes, what is reliable scientific inference based on if not novel prediction?

5 THE POSITIVE THESIS: THE REALIST'S SUCCESS-TO-TRUTH INFERENCE

Our investigation of the negative predictivist thesis has revealed that novel prediction does not appear to have an important epistemic advantage over accommodation due to potential methodological issues with accommodation. However, there is still another, entirely different strategy of defending predictivism: the positive, explanatory argument. According to the positive predictivist thesis, novel prediction provides an advantage because only novel success calls for a truth- or reliability-based explanation. Novel success in the absence of truth or reliability would be 'miraculous,' so we are justified in believing that novelly successful theories contain true theoretical components.

Based on the findings of the previous chapter, the positive thesis now also faces new challenges. We found no indication that novelly successful theories are generally more likely to be true than accommodatively successful theories. The only contexts where novel prediction provides some advantages are contexts where our epistemic situation is weak to begin with, and the benefits of novel prediction are limited to providing modest inductive support. However, importantly, we never *explained* how novelly successful theories emerge in the absence of truth or reliability. The discussion was also focused chiefly on statistical hypothesis testing and on fields of science that some realists might not consider 'mature' enough for realist commitment, given the ongoing methodological crisis (see Psillos 1999, p. 102; Chakravartty 2017a). To evaluate the predictivist position in full, we need to also consider the positive predictivist argument on its own terms, as pursued in the context of the scientific realism debate.

The novelty-based defense of scientific realism, as we can recall, appeals to surprising novel predictive success as the criterion for what warrants realist commitment in science. Per the divide et impera strategy, the realist commits to relevant theoretical constituents (e.g. unobservable theoretical mechanisms, laws, and entities that feature in scientific theories) rather than theories as a whole. The argument that justifies the inference from surprising novel success to theoretical truth is the NMA, according to which truth or reliability provides the best explanation for novel success in science. There are different ways to fill the details of the argument, but in the basic form it either appeals to the reliability of scientific methodology in generating true theoretical constituents (e.g. Psillos 1999), or to the strong intuition that true theoretical constituents are the best explanation for impressive novel success (e.g. Vickers 2019; see also Worrall 2011).

The purpose of this chapter is to evaluate the novelty-based realist's NMA. To do this, we first need to become clearer of the explanatory landscape in this context: what could, if not true theoretical components (or reliable selection of true theoretical components), explain the novel successes

that we see in science? This is discussed in section 5.1, which introduces and clarifies van Fraassen's (1980) alternative anti-realist explanation. With an understanding of the potential alternatives for explaining novel success, section 5.2 proceeds to consider the anti-realist's alternative more closely. I argue that contrary to the realist's assumptions, theoretical truth does not constitute a unique or best explanation for novel success in science. This is due to what I call novelty-based realism's *contingency problem*: novel success does not dependably implicate truth or reliability in the contingent context of scientific theorizing, where multiple attempts at theorizing are made by multiple scientists. Section 5.3 makes further specifications to the argument and discusses possible objections. Section 5.4 provides an overall conclusion on the issue of novelty-based predictivism.

5.1 ON EXPLANATIONS FOR NOVEL SUCCESS

Chapter 3 introduced various arguments against the novelty-based realist's No Miracles argument. These arguments were deemed insufficient to defeat the NMA and scientific realism; the scientific realist is able to hold on to their preferred position. However, there is one more challenge that addresses more directly the claim that the truth of the theory constitutes the best explanation for novel success in science. This is Bas van Fraassen's selectionist, Darwinian explanation for the success of science. van Fraassen (1980, pp. 39-40) famously compared scientific theorizing to the competition in the state of nature, writing that scientific theories are "born into a life of fierce competition, a jungle red in tooth and claw." He argues that in such an environment, the success of (current) scientific theories is no miracle that stands in need of realist explanation: the fact that scientists actively select empirically successful theories and discard all empirically unsuccessful theories explains why empirically successful theories survive in the scientific process.

As it stands, scientific realists have not considered van Fraassen's selectionist response very compelling (see, for example, Musgrave 1988, p. 242; Psillos 1999, pp. 93-94; Psillos 2020, p. 23). Musgrave (ibid.) was the first to give what has become the standard retort: van Fraassen changes the subject. van Fraassen highlights one type of explanation for why successful theories survive in the scientific process, one that realists too can endorse: scientists actively select successful rather than unsuccessful theories. However, what this particular explanation does not address is *what it is about a successful theory that makes it successful*. Nature again provides a useful analogy. To explain why an organism is successful biologically, we do not say that if it were not it would have been eliminated. To explain the organism's success, we need to investigate what properties or characteristics of the organism *enable* it to be successful. So, too, in the case of scientific theories. The realist argues that to explain what makes a particular theory capable of achieving important

empirical success such as novel predictive success, we need to attribute some success-enabling property to it, e.g. *truth*. In other words, whereas the realists have been concerned with providing a *genotypic* explanation for the success of scientific theories, van Fraassen has simply provided a *phenotypic* explanation that appeals to a particular theory selection criterion in science (see Psillos 1999, pp. 93-94). The phenotypic explanation is no threat to the realist's genotypic, truth-based explanation for the novel success of scientific theories.

Or, so the argument goes. In this section and the next, I explore more closely the dynamics between the realist's and the anti-realist's competing explanations for the novel successes of science. The realists have argued that van Fraassen changes the subject: instead of discussing the theory, van Fraassen focuses on how scientists select theories in the scientific process. However, we have seen in previous chapters that the realists do also often invoke the theory selection process in formulating their novelty-based NMA (see section 2.5 and Chapter 3). Conversely, van Fraassen does also have an explanation for the success of scientific theories as such: *empirical adequacy*. To be able to evaluate the realist's and the anti-realist's competing explanations, we first need to adequately juxtapose them. There are, in fact, two puzzles here that are under dispute, and both the anti-realist and the realist provide answers to both puzzles. The first puzzle is what explains the novel success of scientific theories as such, i.e. what property should be attributed to the theory itself to explain its success with novel data. The second puzzle is how novelly successful theories are selected in the scientific process. The relationship between these two puzzles, and the realist and the anti-realist solutions to them, are explored in what follows.

Let us start with the explanations for the novel success of scientific theories as such. The realist argues that the best explanation for the novel success of scientific theories are their true theoretical constituents, while the anti-realist argues that all we need to believe is that the theory at hand is empirically adequate, whether or not novel prediction were made. How are *these* competing explanations? Here, we should first recognize that there is important agreement between the realist and the anti-realist. Both the realist and the anti-realist believe in what science says about the observable world (see section 3.1). Both agree that there are observable regularities in nature, and that scientists actively look for such regularities and theorize about them. van Fraassen (1980, 40) ends his selectionist argument by emphasizing this, writing about the scientific process: "[o]nly the successful theories survive – the ones which *in fact* latched on to actual regularities in nature." The dispute between the realist and the anti-realist is targeted specifically on the *un*observable laws, mechanisms, and entities that scientific theories posit to explain observable findings and regularities. The anti-realist argues that unobservable theoretical posits *could be* false despite being associated with novel successes in science. The realist, in contrast, thinks that belief in the truth of such posits is epistemically justified (based on the NMA).

From this dialectic, we get two alternative explanations for the novel success of theories as such in science. To explain novel success, we can either hold that we have in hand a) a theory that contains *both* true observable and true unobservable constituents or b) a theory that contains true observable but false unobservable constituents. The realist endorses a, whereas the anti-realist argues that a is not justifiably superior to b, so we should suspend judgment about the question between a and b. In other words, the realist recommends that we choose one of these options (a), but the anti-realist thinks the choice between them is underdetermined in our epistemic situation.

This has been thought problematic to the anti-realist. For how does alternative b actually *explain* the success of the theory? How could false unobservable constituents feature in an equally compelling explanation for the success of the theory with novel data as true unobservable constituents? Realists have used terms such as 'chance,' 'coincidence,' and 'miracle' (e.g. Musgrave 1988; Psillos 1999; Barnes 2008) to describe the anti-realist position. The realists argue that anti-realists are committed to empirical success being due to some kind of miracle, whereas the realists are able to hold that success is due to something concrete: theoretical truth. However, this does not appear to be a fair characterization of the anti-realist position. As we have seen, the anti-realists do not argue that the scientific theory is successful due to chance. Rather, the anti-realist argues that the theory is successful *because it has latched on to actual regularities in nature* (see van Fraassen 1980, p. 39-40). In other words, the anti-realist thinks that the theory is successful not by accident but because it has latched on to something real. The difference is that the anti-realist thinks that certain *parts* of the theory need not be true despite its success in latching on to observable reality (i.e. the unobservable parts). The alternative explanation that the anti-realist puts forward is not about chance but about *epistemic modesty*: we do not need to appeal to extra theoretical truth to explain the success of scientific theories with observable regularities in nature.

But, why should we not? How could *explanation b* be equally compelling (or perhaps better) to *explanation a* in cases of novel success? This, I suggest, is where the discussion turns to the second problem: how do scientists select theoretical constituents in the pursuit of novel successes? The realist presents one answer: scientists select theoretical constituents into their novelly successful theories *reliably*, i.e. they do it in such a way that they end up with true rather than false theoretical constituents in their novelly successful theories (e.g. Psillos 1999; White 2003; Barnes 2008). The anti-realist has, again, been argued to be committed to the idea that the theory generation or selection process in science is based on 'chance' or 'coincidence:' supposedly, scientists select theories *either* reliably or by chance (see, in particular, Barnes 2008). But, this remains too simplistic. First, as argued already by Maher (1990), we are not faced with a dichotomous choice between chance and reliability when we seek to explain how successful theories emerge

in science. The reality lies somewhere in between: scientists are more or less reliable in generating and selecting true theoretical components. Second, the anti-realists are certainly not committed to the idea that scientific theorizing is anything like a random chance process where scientists churn out random strings of text and mathematical formulas and apply them to observations. The anti-realist holds that science proceeds very much *non*-randomly: the purpose of the enterprise is to gather empirical evidence, look for observable regularities, and theorize about those regularities. All that the anti-realist may believe in contrast to the realist is that the process of theorizing about unobservables need not be all that reliable: despite the empirical success science has reached with the observable world, scientists may not be very reliable with regard to the unobservable world.[64]

Why does the anti-realist hold this position? This, I believe, is where the selectionist response becomes salient and is applied in the discussion. The realist argues that the best explanation for the emergence of novel success is the reliable selection of true theoretical constituents into the theory. The anti-realist, following van Fraassen, responds something like follows. Scientists achieve novel successes because scientists keep on theorizing until they, in fact, latch on to an actual novel finding or regularity in nature; when that happens, they keep the theory, and in cases where this fails to happen, they discard the theory. In other words, scientists actively look for novel success, and continue to try until they achieve it. No further explanation, reliability-based or otherwise, for how novel successes are found in science is needed: scientist look for novel success, and they select the theories that achieve it.

The purpose of the selectionist response, and where it becomes relevant to the discussion, is now perhaps starting to become clearer. The role of the selectionist argument is to highlight certain points about how scientific theories are selected, which is *then* meant to establish that there is no need for a reliability-based explanation, and thus no need to attribute truth to novelly successful theories. However, there is a sense in which this response is yet incomplete, and the realists are correct in insisting that van Fraassen's argument does not defeat the NMA. The observation that scientists actively select for novel success rather than failure does not undermine, as such, the claim that *the best explanation* for novel predictive success is the reliable selection of true theoretical components into the theory. Even if we agree that scientists keep on pursuing novel success until it is eventually achieved, it is not clear why we should not believe that scientists are reliable in the instances where it is ultimately achieved. And, this still leaves the realist explanation for the novel success of scientific theories, i.e. *explanation a*, on the table as a viable option. If the selectionist challenge is to work, it needs to show *why* the kind of theory generation and selection mechanism that operates in science

[64] Recall Wray (2018, p. 71): "I think that even the most skeptical contemporary anti-realists would think that it is doubtful that all the theoretical entities postulated more than fifty years ago do not exist."

means that reliable selection of theoretical truth need not be invoked to explain the novel successes of science. If the anti-realist is able to do this, then *explanation a* and *explanation b* are shown equally viable. In the next section, I argue that the selectionist response, if developed more fully, does precisely this.

To recap, with the NMA and the selectionist challenge to it, there are two problems to be explained and two explanations for both problems. The first problem is why a particular novelly successful theory is successful with the novel data. The realist argues that the best explanation invokes true theoretical components while the anti-realist argues that all that we need to believe is that the theory has latched on to actual regularities in nature (i.e. it is empirically adequate). The second problem is how the scientists produced or selected such a theory. The realist argues that the scientists reliably selected true theoretical constituents into the theory while the anti-realist holds that reliability is not required. The selectionist challenge to the NMA arises from the anti-realist's answer to the second problem. If there is something about the scientific process that allows for novel successes to arise without the need to invoke the (reliable) selection of true theoretical components into the novelly successful theory, true theoretical components are not needed to explain the novel success of scientific theories; empirical adequacy will do.[65]

5.2 THE SELECTIONIST RESPONSE AND THE CONTINGENCY PROBLEM

I now proceed to develop the selectionist challenge further. I suggest that this challenge works in particular against the claim that *novel* success is a dependable indicator of theoretical truth in science. To see why this is so, we need to explore the selectionist response more closely, and consider in what way it challenges the realist's novelty-based NMA as *an abductive, explanatory* argument. I will start with the overall argument, and then move to consider a case example from science that illustrates the argument in practice. In section, 5.3, different aspects of the new, updated version of the selectionist argument are further explored and clarified.

The selectionist response, as originally formulated by van Fraassen, sought to challenge the NMA by appealing to the theory selection criterion that scientists use in scientific practice. However, I believe that the strength of this response lies not in the observation that scientists actively select for novel success rather than failure, but rather in what van Fraassen's

[65] A similar argument has also recently been made by Boyce (2018). He distinguishes the truth-based explanation for empirical success as 'the miraculous theory argument' and the selection-based explanation as 'the miraculous choice argument' (following Barnes 2002b), and argues that making the selection-based argument is enough for the anti-realist's purposes. The success of our best scientific theories can then be 'an extraordinary coincidence' (Boyce 2018). I argue below that the purpose of the selectionist argument is to establish the opposite: the novel success of scientific theories need *not* be an extraordinary coincidence even if their unobservable constituents are false.

argument implies about the reliability or unreliability of the theory generation or selection process in science. The underlying challenge to the NMA arises from particular features of the scientific process that van Fraassen tacitly invokes in his selectionist challenge, when he refers to the fierce competition that scientific theories face. These are a) *the multiple attempts* that scientists make while theorizing about the world, and b) *the interdependent* nature of the scientific process of theory generation. These factors, I suggest, act together in a way that undermines the reliability-based explanation (and thus the truth-based explanation) for novel successes in science. The emergence of novel successes in science can be explained by *the trial and error process* of scientific theorizing, which does not require or implicate any particular degree of reliability in cases of novel success. Rather, the relationship between novel predictive success and the veridicality of the theory from which it is derived can be properly seen as a contingent accident.

We can start by illustrating the argument through a simple example. This example is, in fact, already familiar in the prediction versus accommodation literature, and the gist of what I am about to argue has been recognized previously, at least in part, by others (see, in particular, Barnes 1996; see also White 2003). Novel predictive success, as we can recall, is supposed to require a reliability- or truth-based explanation because the alternative is 'a miracle.' However, in certain contexts, there is a rather mundane explanation for novel success that does not require either truth, reliability, or miracles. This explanation is: many people tried many times until somebody happened to produce something surprising (cf. van Fraassen 1980, pp. 39-40). Barnes (1996) demonstrates this with the example of coin flipping.[66] Consider that somebody has predicted the outcome of a coin flip ten times in a row – i.e. she has achieved highly surprising novel success. This generates a strong predictivist intuition that the predictor must have some reliable insight into "the coin flipping process" (see Maher 1988; Worrall 2014). However, suppose that we then learn that the predictor comes from a large community of coin flippers where many people tried many times to predict ten in a row. In such an environment, the fact that somebody sometimes gets ten in a row does not require a reliability-based explanation, as long as we know that enough have been attempts made: with enough attempts, we can easily calculate that such successes are likely to arise by chance (see Barnes 1996 for such calculations). In other words, when the overall, community-level process for the attempts at coin flip prediction is taken into account, reliability no longer constitutes a unique or best explanation for novel success in this context: multiple (unreliable) attempts do the job.

[66] Barnes himself did not view this as a problem that challenges scientific realism, arguing that the number of scientist working on theories within a particular domain is typically small (see Barnes 1996, 2008). I argue below that Barnes has not recognized the full extent of the challenge that this poses to the novelty-based realist argument.

Science is not like coin flipping where scientists randomly generate novel ideas with an ostensible probability of success (see Howson & Franklin 1991; Harker 2006). Accordingly, we should resist drawing a direct analogy between this example and scientific practice (cf. Barnes 2008; see also section 5.3).[67] However, I suggest that the principle that the community-level process for the generation of novel ideas influences the explanation that is called for in cases of surprising novel success applies in science as well. In fact, I believe that it applies particularly strongly in science, due to fundamental features of the scientific process. The scientific process by its very nature is geared towards generating heterogeneous theoretical environments where, I argue, there is no need to invoke truth- or reliability-based explanations for novel success:

First, in science, if anywhere, the assumption that novel successes are pursued through *multiple attempts* is met (similar to the coin flip example).[68] Science is a practice that involves a very large community of people, coming from multiple generations and all regions of the world. In the scientific process, scientists generate vast numbers of theories and hypotheses over a very long period of time, which are gradually tested against a growing base of empirical evidence. When theories fail, scientists come up with new alternatives, which then have their turn with observations and experiments. This process repeats in multiple iterations, sometimes with just one or two pieces of evidence that act as the catalysts for new ideas, as scientists gradually whittle down the alternatives about what could represent the truth about the world. All in all, a *very* large number of attempts end up being made in the process of scientific discovery towards important new findings.

Second, unlike in the coin flipping example, the many attempts that scientists make at developing new theories are *interdependent*. When a scientist proposes a new theory, it will invariably depend on evidence gathered by other scientists, previous ideas generated by other scientists, and the possibilities that are open for exploration in light of that previous knowledge and evidence. Scientists rely each other's work in both collaborative and competitive ways. In terms of collaboration, they use ideas generated by others to introduce novel tweaks and adjustments. In terms of competition, they study the prevailing theoretical environment and attempt to generate ideas that others have not yet thought of. In this kind of process, large and heterogeneous pools of theoretical constituents and sources of evidence are created that the scientists gradually develop and explore on their way to novel discovery.

[67] Statistical hypothesis testing is an exception. These type of cases were considered in Chapter 4. This chapter expands the perspective by focusing on the explanatory NMA.

[68] See Fahrbach (2009) for more detailed calculations about the volume of scientific work in the world. For example, he calculates that the number of published scientific research articles between the year 1600 and 2010 is around 60-80 million. These articles should by no means all be thought to contain novel ideas, but the figure gives some indication of the amount of work that has been conducted in the scientific process. Fahrbach further calculates that the doubling rate of publications has been about 15-22 years for the past 300 years.

Consider, then, the dynamics between the realist and the anti-realist explanations for the emergence of novelly successful theories in science. Both the realist and the anti-realist agree that scientists collect empirical evidence, look for observable facts and regularities in the world, and attempt to theorize about those facts and regularities. Both agree that scientists are successful with regard to the observable world and have discovered new observable facts and regularities in the scientific process. In terms of theoretical truth (or the reliability of the selection process for theoretical constituents), what is the best explanation for novel success in this kind of a context where numerous, mutually interdependent attempts at theorizing are made over a very long period of time? I submit that, contrary to the realist's assumption, no best or unique explanation can be expected. Instead, the answer is going to be 'it depends.' In an environment where so many attempts are made over such a long period, any number of explanations for how particular novel successes are achieved are likely be found. All it takes, as the anti-realist observes, is that occasionally somebody hits on an actual regularity in the world without invoking it in the process of theorizing (i.e. they achieve use-novel success). The fact that the pattern or regularity is real explains why the prediction itself is successful. However, as the new theory emerged in a process where many attempts were made, true theoretical constituents need not be invoked to explain that success; the many attempts made it possible (and ultimately likely) that somebody was going to hit on a novel pattern in the world regardless of the veridicality of their theory. No particular degree of reliability in the selection process or amount of true theoretical constituents in the theory need be invoked to explain that success.

The relevant dynamics can be illustrated perhaps most vividly through a case example. An interesting example is provided by the continental drift controversy, or the 'mobilism' vs. 'fixism' controversy, which focused on one of the most important geological problems in recent scientific history: is there horizontal displacement of the continents (see Frankel 2012a-d)? This controversy unfolded over multiple decades during the 20th century, until it was settled in favor of continental drift and plate tectonics in the 1960s. The topic has recently been made accessible to the philosophical audience by the philosopher Henry Frankel, whose decades-long investigation into the contours of this debate culminated in his four-volume work Frankel (2012a-d). Frankel focuses on not only the successes of the scientific process, but also the myriad of false theories and hypotheses that were proposed on the way to those successes, providing a unique perspective into the inner workings of the scientific process.

The particular case of scientific success that is of interest to us here is Harry Hess's mantle-convection based theory of continental drift and seafloor spreading. Frankel (2012d) shows that continental drift triumphed over fixism after the confirmation of Hess's new seafloor spreading hypothesis in the 1960s. Hess (1962) proposed an innovative novel theory about convection in the Earth's mantle, according to which convection currents

cause seafloor to spread along mid-oceanic ridges, the long ridges that stretch thousands of miles along the ocean floor. Hess postulated that in this process of mantle convection, the seafloor gradually spreads, becoming wider and wider, and continental drift is propelled as the continents ride passively on the mantle with convection.

Frankel (2012c, p. 320) explains that at the time, Hess's theory was considered highly controversial, as there was no compelling reason to believe that either seafloor spreading or continental drift happens. However, Hess's theory led to some very impressive (temporally and use-) novel successes, as impressive as any that we have seen in the history of science. In particular, Vine and Matthews (1963) showed that the seafloor spreading hypothesis, if connected with paleomagnetic evidence, implied that magnetic surveys of the ocean floor should show an alternating 'zebra' pattern of normally and reversely magnetized material that runs parallel to oceanic ridges – a 'recording' of sorts of the gradual creation and spread of new seafloor. This highly surprising and controversial prediction was spectacularly confirmed, leading in short time to near universal acceptance of seafloor spreading and continental drift – even among many diehard opponents of drift (see Frankel 2012d).

The novelty-based NMA implies that given the highly surprising novel success of Hess's theory, it should be likely that the theoretical apparatus Hess put forward contains important true constituents. However, this is not what the scientific literature indicates is the case. The current leading theory on the principal cause of continental drift and seafloor spreading is a 'top down' theory, according to which the plates themselves are responsible for the buoyancy forces that drive their motion (plate movement happens primarily in consequence of 'slab pull' and 'ridge push') (see Forte 2011; Rowley et al. 2016).[69] In fact, the confirmatory status of Hess's mantle-focused theory was considered uncertain even after the confirmation of seafloor spreading and continental drift, and the question of the driving forces of plate movement was still very much an open problem in geology (see Forsyth and Uyeda 1975; Frankel 2012d). In other words, despite its highly impressive novel success, Hess's theory was not a) perceived strongly confirmed by its novel predictions, nor b) is it a theory that captures the truth about the forces that drive plate movement.

How can such impressive and surprising novel success and the truth of the theoretical mechanism from which it is derived diverge in such a marked way in science? The answer, I suggest, is that when we consider cases of surprising novel success in science, we cannot just focus on a single isolated theory but must see the theory as a part of the theoretical environment of its time. Surprising novel success can very well emerge without a reliable connection to theoretical truth as long as scientists a) make many attempts at

[69] The driving forces of plate movement are still debated in geology (see Rowley et al. 2016 for a challenge to the received top down view).

developing new theories, and b) in the process they populate the prevailing theoretical environment with multiple theoretical constituents (some of which may be false and others true). Multiple attempts in a heterogeneous theoretical environment make novel success possible with just about any combination of true and false constituents. In the case of Hess's convection-based theory of continental drift and seafloor spreading, this theory emerged from a long line of more or less (un)successful theories and hypotheses proposed on geological phenomena over multiple decades. To get a sense of the range of ideas that were put forward about geological phenomena prior to and around the time of Hess's success, consider the following non-exhaustive, roughly chronological list of theories, discussed in Frankel (2012a-d).

Theory	Sample claim(s)
Suess's contractionism (a, pp. 39-42)	- The Earth has always been cooling and contracting - Present-day oceans were formed by sinking of parts of former continents - Horizontal thrusting and folding produces mountain systems and island arcs
Wegener's continental drift (a, pp. 50-61, 159-162)	- Continents plow through the ocean floor - Continental drift is caused by centrifugal and tidal forces - Mountains are formed at the leading edge of drifting continents - Island arcs are slivers broken off at the trailing edge of drifting continents
Taylor's continental creep (a, pp. 65-69)	- A tidal pull created by Earth's capture of the Moon caused the continents to creep towards the Equator - Mountains and island arcs are formed in consequence of compression at the leading edge of moving continents
Landbridges (fixism) (a, pp. 92-98)	- Trans-Pacific landbridges once existed connecting the unmoving continents
Isthmian links (fixism) (a, pp. 107-112)	- Isthmian links (i.e. narrow strips of land) once connected the unmoving continents

Daly's downsliding (a, pp. 172-179)	- Continents slide down and sink into the hotter substratum (i.e. mantle) due to gravity - Mountains are elevated due to thermal expansion at their roots - New ocean basins form in the spaces created in the lee of sliding continents
Joly's thermal cycles (a, pp. 183-190)	- Continents sink and rise periodically due to heat changes in the Earth's interior
Holmes's substratum (i.e. mantle) convection (a, pp. 210-223)	- Convection currents in the substratum (i.e. mantle), fueled by radiogenic heat, move continents - Continents override the ocean floor - Mountain belts rise out of interior geosynclines
Vening Meinezs's fixist mantle convection (a, pp. 231-232)	- Convection currents cause the formation of island arcs and oceanic trenches, while the continents remain fixed in place
Griggs's fixist theory of mountain building (a, pp. 234-238)	- Convection cycles in the mantle beneath unmoving continents produce mountain-building cycles
Egyed's slow Earth expansion (b, pp. 279-290)	- The Earth's core contains matter in an unstable, high-pressure state, which leads to gradual increase of the Earth's volume - The continents drift apart as the Earth expands
Carey's accelerating Earth expansion (b, pp. 325-330)	- A phase change in Earth's interior can produce a forty-fold increase in Earth's volume - New seafloor is created at oceanic ridges as the Earth dilates, increasing the separation of the continents
Heezen's rapid Earth expansion (c, 400-408)	- The Earth expands in consequence of density changes within Earth's interior and the increase of the gravity constant G over time - Ocean basins evolve from rift valleys as the Earth expands

Runcorn's convection-based theory(/ies) of drift (c, pp. 18-25)	- Earth's growing core creates convection cells that move continents
Hess's theories on the origin of oceanic ridges	- The Mid-Atlantic Ridge is an old folded mountain system (c, p. 209) - The Mid-Atlantic Ridge was caused by a rising mantle convection current and the ridge subsided when the convection stopped (c, pp. 224-225) - Rising mantle convection causes ridges to subside (c, pp. 227-229) - Mantle convection currents propel seafloor spreading and continental drift along oceanic ridges (c, pp. 243-254)
Ewing and Ewing's fixist theory of the origin of oceanic ridges (c, pp. 408-410)	- Convection deep in the mantle creates an upwelling of basalt magma at oceanic ridges - Continents remain fixed in place
Menard's theory of the origin of oceanic ridges (c, pp. 333-337)	- Mantle convection currents cause elevation and depression of oceanic ridges
Menard's seafloor stretching (c, pp. 348-352)	- Mantle convection currents stretch seafloor along oceanic ridges
"The Lamont group's" tension theory of oceanic trenches (c, pp. 259-261)	- Oceanic trenches form in consequence of tensional forces
Officer's (et al.) compression theory of oceanic trenches (c, pp. 263-265)	- Oceanic trenches from in consequence of horizontal compression
Hess's and Fisher's convection theory of oceanic trenches (c, pp. 265-271)	- Oceanic trenches form in consequence of downward convection

Table 1. A sample list of theories within the continental drift controversy. The references are to Frankel (2012a-c).

The theories and hypotheses listed in Table 1 illustrate the scope and the type of ideas that emerge in the process of scientific theorizing. The continents, for example, were hypothesized to be subject to both contraction and dilation, to sink and to rise, and to move and to stay fixed, and multiple theories and hypotheses were developed about the mechanism behind each possibility. Similarly, the ocean floors were thought to be permanent, to sink, to stretch,

to expand, to carry the drifting continents, etc. Almost all of these theories and hypotheses are (at least for the most part) false.[70]

Hess's novelly successful theory, appearing towards the end of the list, arose in consequence of Hess's multiple attempts to explain the origin of mid-oceanic ridges. He connected their origin to the idea that there are convection currents in the mantle – a speculative idea that was applied to various geological problems at the time (see Table 1) – and after many attempts postulated that mantle convection causes the seafloor to spread. When we consider how likely it is that a theory like this is made up of true constituents, I suggest that this is a matter of deep historical contingency in the scientific context. In the process of scientific theorizing, scientists go through a trial and error process, where they populate the theoretical environment with multiple theoretical constituents, which are combined in various ways in different theories. When scientists make new attempts at theorizing, novel successes can arise with any combinations of true or false constituents in the theory. In this case, Hess ingenuously came up with a novel idea about a surprising new phenomenon, but just happened to use false theoretical constituents in his theory.[71]

The process through which important novel ideas emerge in the absence of reliability can also be illustrated by the case of Hess the individual. Frankel (2012c, p. 201-202) discusses Hess's method to science, emphasizing his courage to explore novel hypotheses. Hess himself explained that novel but incorrect hypotheses were "an invaluable often necessary stepping stone to a better hypothesis" (see Frankel 2012c, p. 202). He believed in the value of introducing wrong hypotheses into the theoretical environment so that they can act as catalysts to better ideas and was not afraid to develop many such hypotheses himself. Hess repeatedly tried to explain the origin of island arcs, mountains, oceanic trenches, ridges, and crust, generating multiple false theories and hypotheses, until he finally produced the novel (and correct) idea of seafloor spreading. Hess's attitude and efforts as a scientists illustrate on a smaller scale the process that I suggest challenges the realist's explanation for novel success: if you try many times (using new empirical evidence as it comes in), there is a chance that at some point some new idea is going to latch on to an actual regularity or pattern in the world. Crucially, reliability is not needed to explain how that happens; trial and error will do.[72]

---

[70] The list in Table 1 is by no means exhaustive. Missing are yet numerous other (false) theories and hypotheses (e.g. the fixist works of Jeffreys, Tuzo Wilson's multiple hypotheses on both the mobilist and fixist sides), alternative or tweaked versions of these and other theories and hypotheses, and more local disputes about geological phenomena and evidence all over the world.

[71] Other possibilities were available in the prevailing theoretical environment. For example, the seafloor spreading hypothesis is also a natural consequence of Carey's Earth expansion theory, which Hess may have been aware of (see Frankel 2012b, pp. 329-330). However, Hess had already previously rejected Earth expansion (ibid.). A theory that arguably comes closer to modern views, Daly's downsliding, had also been introduced into the literature decades earlier, but at the time it was considered questionable even by Daly himself (see Frankel 2012a, p. 178, 181).

[72] We can now briefly return to the open issue of agent-based predictivism. Trial and error, other than the theorist's reliability, can explain the novel successes that we in science (see section 2.5). In the

At this point, the realist may step back and object: we have not yet taken into account all of the constituents of the novelty-based NMA. Namely, we have yet to consider *degree of surprisingness* as a relevant factor to the argument. The realist argues that realism becomes more and more plausible as the surprisingness of the novel findings increases (see Vickers 2013, 2019). What ought we to think about this specification? Does increasing the surprisingness of the predicted empirical results help with the contingency issue that we have raised here?

The answer, I suggest, is 'no:' now we are simply introducing more contingency into the mix. The degree to which novel successes appear surprising or impressive is itself influenced by any number of historical contingencies related to the contemporary evidence base, prevailing theories, opinions of leading scientists (etc.), which are independent of the veridicality of the theory at hand. We can again take Hess's case as an example. Based on Frankel (2012a-d), a number of factors contributed to making the novel success of seafloor spreading and continental drift more surprising. For example, 1) continental drift was in disrepute due to the perception that Alfred Wegener's original theory of drift had been significantly flawed, and 2) continental drift was fiercely opposed by one of the leading geophysicists of the 20th century, Harold Jeffreys (e.g. Frankel 2012c, pp. 137-143). It is a simple exercise to consider alternative ways in which history could have unfolded where we hold other things equal but vary such factors. For example, continental drift might have been seen as a much more likely contender if the idea had been advocated by somebody like Jeffreys. This would have contributed to making Hess's novel success with seafloor spreading less surprising, as continental movement had already been expected as a likely possibility. On the other hand, it is also possible to think of ways in which Hess's success could have been even more surprising. Say, for example, that some misleading evidence had existed that (unknown to Hess) indicated that the seafloor was uniformly old. The success of the seafloor spreading hypothesis would then have been even more striking. However, no matter which way we turn the dial, increasing or decreasing surprisingness, it is hard to see how this could influence how likely it is that the mantle-based theory that Hess proposed is true (cf. Collins 1994).[73]

contingent context of scientific theorizing, novel success does not provide reliable evidence about the degree of reliability of the theorist, so there is no dependable evidence to gain about the theory based on what novel success implies or does not imply about the theorist.

[73] It is also possible to think of other contingent factors that simultaneously influence what kind of theory Hess – or somebody else – is going to propose (e.g. somebody in our counterfactual world discovers some new evidence that Hess is then made aware of). With these factors as well, the relationship between surprisingness and truth remains circumstantial, and a case could even be made that the surprisingness of an empirical fact decreases the likelihood that somebody is going to generate a true theory about it. For example, if a phenomenon such as seafloor spreading becomes highly surprising through misleading evidence which suggests that the seafloor is uniformly old, this might prevent theorists such as Hess from theorizing about it to begin with. But, if the phenomenon is already expected in the theoretical environment, it should be more likely that somebody will be able to produce a veridical theory about it.

To sum up, I have argued that, contrary to what the realist claims, there is no best or unique explanation that is implicated in the case of surprising novel success in science. Given the multiple interdependent attempts that scientists make at theorizing about empirical facts and regularities, generating highly heterogeneous theoretical environments from which theorists draw new ideas, there is no reason to invoke reliable selection of true theoretical components in particular as the explanation for contingent novel successes. In this kind of a competitive environment, novel successes can arise with any combinations of true or false constituents in the theory. To be able to explain surprising novel success in science, all we need is to hold that there is an actual regularity in the world, as the anti-realists argue. Moving beyond this to commit to unobservable theoretical constituents remains unjustified, at least going by the abductive predictivist basis (i.e. the novelty-based NMA).

Our discussion has further revealed the extent to which the novelty-based realist is dependent on continent accidents, a point with which we began this study (see Chapter 1). We have seen that in order for the novelty-based realist's NMA to land successfully, a number of historical accidents must take place: 1) a theorist must be in a position to develop a new theory using true theoretical constituents, while 2) not using some empirical results, whose 3) prediction is considered surprising. In the scientific process where multiple attempts at theorizing are made over a long period of time and the views on what is considered surprising are at all times evolving based on the contemporary evidence base and theoretical ideas, there is little reason to trust that such contingent accidents should co-occur in any kind of dependable way. In accordance, neither truth nor reliability need be invoked to explain the novel success of particular scientific theories, and neither novel predictive success nor accommodative success calls for a realist explanation.

5.3 SPECIFICATIONS AND OBJECTIONS

We should next evaluate the argument that I have produced more closely, and discuss potential objections. First, let us take a closer look at the particular case example I have used. Does the case of Hess and his seafloor spreading hypothesis meet the criteria that the realists have set for their novelty-based NMA (i.e. surprisingness, selective realism)?

I suggest that the example I have used ticks all the boxes. First of all, in this case if anywhere in science we have *surprising* novel success. Continental drift was fiercely opposed by prominent geologists and geophysicists at the time when seafloor spreading was confirmed, many of whom considered the idea of drift highly improbable (see Frankel 2012d). Seafloor spreading was highly surprising even to the inventor of the Vine-Matthews hypothesis, Frederick Vine, who did not think the hypothesis had much of a chance of being published (see Frankel 2012d, pp. 139-140). (In fact,

the same hypothesis was also proposed by another geophysicist, Lawrence Morley, whose paper was rejected in the peer review process at the time [see Frankel 2012d].) The novel confirmation of the seafloor spreading hypothesis certainly counts as surprising success in scientific practice, if anything does. Second, it also seems to me that this example is difficult to resist based on the accounts of selective realism that have been defended aligned with the novelty-based NMA (e.g. Psillos 1999; Vickers 2013). A selective realist could perhaps seek to argue that the theoretical mechanism Hess introduced to account for seafloor spreading and continental drift is not necessary to derive the novel success of the seafloor spreading hypothesis, so it can be eliminated as an idle posit (the existence of seafloor spreading, after all, will do). However, the problem with this response is that it is difficult to see how the realist can then avoid the *overexclusion* of theoretical posits. If the selective realist uses a loose enough criterion that allows the entire (unobservable) theoretical mechanism postulated by the theorist to be eliminated, there is not much left over what the anti-realist does not already accept. Furthermore, even if this strategy somehow succeeds, it does not touch my main argument, which targets the contingent and undependable relationship between novel success and truth. The anti-realist would be very happy if the realist is willing to concede that just about all of the theoretical mechanism can very well be eliminated even if we are talking about a case of highly surprising novel success.

Another challenge that targets the specific example of Hess proceeds as follows. Scientists did establish something important with Hess's novel success: seafloor spreading and continental drift. Is this not enough to support realism about important constituents of Hess's theory? The realists have been after the confirmation of important unobservable theoretical mechanisms, laws, and entities (see Psillos 1999, pp. xx-xxi), and my main concern here has been to target this claim. The continents and the seafloor are both observable features of the natural world, and the discovery that the seafloor spreads so that the continents move is certainly an important finding that concerns real phenomena in the world. However, I take it that the central issue in the realism debate concerns *the theory* about these phenomena, which in this case was wrong. If all that the realist believes over the anti-realist is that certain observable parts of the world go through changes that are not immediately visible to the naked eye, we have come far from the substantive realist beliefs about unobservable theoretical mechanisms, laws, and entities. Furthermore, it is also unclear to me if even (at least some) anti-realists have a reason to doubt the existence of these particular phenomena. The seafloor spreads in many locations by several centimeters per year, which is something that may be observable.[74] The evidence about these phenomena being real also

[74] Scientists did actually send a manned submersible to make direct observations of a seafloor spreading center, less than a decade after the confirmation of the seafloor spreading hypothesis (see Heirtzler & Le Pichon 1974). Long term observation of seafloor spreading appears possible in principle in a similar way as long term observation of other relatively slow but observable processes such as glacier size changes.

appears particularly exhaustive of relevant alternatives. For example, common technologies such as GPS have to constantly correct for continental movement. Seafloor spreading and continental drift are among the less controversial findings of science, perhaps more in a category which includes things like microscopic entities and dinosaurs rather than the category of unobservable laws and mechanisms that govern the observable regularities in the world (cf. Stanford 2006, pp. 33-34).

Other potential lines of critique target my argument more generally. First, let us consider generalizability. I have argued that the relationship between surprising novel success and truth is deeply contingent in science, but what ought we to believe about the *correlation* between them? Could it be that despite it being a matter of deep contingency, surprising novel success and truth *are* actually correlated in scientific practice (as it has unfolded over the centuries in the actual world)?

My reply is twofold. First, switching to a discussion about correlation changes the issue, and misses the purpose of my argument. The argument is a response to the realist's *abductive* NMA. The realist has argued that truth (or reliable selection of truth) constitutes the best or the only explanation for novel success in science. I have shown that there is no particular need to invoke truth, or any number of true theoretical constituents, to explain novel successes. The purpose of this response is to challenge the realist's argument on its own terms rather than expand the Pessimistic Meta-Induction strategy. As far as the argument I have presented is successful, it acts as a defeater to the realist's abductive NMA: there is no need to invoke truth or reliability to explain novel success in science.

Second, even if correlation is not the chief issue here, we *should* be interested in any evidence there is about the relevant correlation. (I would certainly count it as evidence *for* novelty-based realism if the realists can demonstrate a strong correlation between novelty and what are taken as true theoretical posits in science.) However, the problem for the realist is that the evidence that we have does not look promising. We have already seen a number of examples of novel success apparently arising from false theoretical constituents (see sections 3.3 and 5.2). But, the example of continental drift reveals an even more fundamental issue, which indicates that these counterexamples may multiply rather easily. Hess's case demonstrates a simple way in which theorists can achieve novel predictive success in science: theorize about an unconfirmed or undiscovered (real) regularity; when the regularity is confirmed by independent experiments, novel success is achieved. This kind of novel success is not in any sense rare in science. To see many more examples of such success, we do not have to leave the continental drift controversy. Just about all the theorists who advocated for various theories about continental movement prior to plate tectonics have achieved novel success in the view of later science, as continental drift has now been confirmed by independent means (see Table 1). Generalizing this observation

to science at large is likely to wreak havoc on the idea that novel success is likely positively correlated with the truth (cf. Lyons 2002, 2006).

Here is another way to look at this issue. There is a simple, general reason for why we should expect novel success to be *negatively* correlated with the truth: scientists generate a vast amount of new ideas in an environment where truth is bound to be relatively rare. After all, on any given scientific problem, there can only be one, or at most a few, theories that really are predominantly true. In contrast, there may be many more false theories that imply some empirical results related to that problem (consider, for example, all the false theories that imply continental drift). In cases where novel success is achieved, it is arguably more likely that the success arises from this larger set of false theories than what is bound to be a smaller set of true theories.[75]

A second, general concern proceeds as follows. I have argued that multiple interdependent attempts undermine the reliability- and truth-based explanations for novel success. However, is not the argument contingent on the degree of surprisingness or improbability of the novel finding, so that the more improbable the finding, the more attempts are needed before we can seriously expect such results to arise by chance in the absence of theoretical truth. Coin flipping provides a useful example. Say, we are evaluating predictions about the results of 5 coin tosses. We know that the probability of correctly predicting 5 coin tosses in a row is roughly 0.03 by chance. It is a simple exercise in algebra to calculate how many attempts are needed until we are likely to get that kind of success by chance in multiple trials (in this case, the expected value is 62). Predicting 5 in a row becomes less and less impressive the more trials are included. So, if we have a novel prediction whose prior probability is very low – say, 0.03 – does that not mean that we would need a large number of attempts (e.g. 62) until these attempts begin to undermine the reliability-based explanation?

No. As pointed out by Harker (2006; see also Howson & Franklin 1991), science in general is not in the business of predicting ostensibly random events with a known, objective prior probability.[76] The probabilities that are assigned to scientific hypotheses depend intimately on the theoretical context at the time. This was illustrated by the example of Hess and what was expected

[75] These considerations also show a certain sense in which temporal novelty and theoretical novelty may be quite unreliable indicators of truth. It hardly makes sense to think that the *first* theory to cover some new phenomenon will be considered the all-time most veridical theory about that phenomenon. Scientific theories tend to improve by time. This is illustrated by the case of continental drift, where scientists gradually progressed from simple ideas of continental drift to the modern synthesis of plate tectonics (see Frankel 2012a-d). Another prominent example is the periodic system, whose modern form looks nothing like Mendeleev's early versions that generated impressive novel successes (see Scerri 2007). The one exception may be the general theory of relativity, which did come first, forming the basis of a new understanding of the Universe, but scientists have yet to generate a better alternative.

[76] Statistical hypothesis testing is an exception where we can make calculations about the probability of chance findings, as discussed in the previous chapter. However, we found no evidence that novelly successful theories are more likely to be true than accommodatively successful theories in this context in contemporary scientific practice. On the contrary, surprising novel success was found one of the strongest correlates of unreplicable research (see Open Science Collaboration 2015).

in light of scientific background knowledge in the 1960s. The connection between what is expected and what kind of theoretical constituents (true or false) are available in the theoretical environment is fundamentally a *contingent* one: a low prior probability assigned to a surprising novel consequence does not imply that a vast number of attempts at generating theoretical constituents are needed to offset that probability. Consider seafloor spreading again. Even if a number of prominent scientists happen to be diehard opponents of continental drift, and a very low prior probability is assigned to this hypothesis (say, 0.001), this does not mean that scientists have to generate hundreds of false theoretical components until novel success is expected in combination with false theoretical parts.[77] Rather, what we need is a series of gradual steps where scientists gather more and more empirical evidence and attempt to generate different theories about phenomena and regularities in the world. When novel successes do arise, the degree of truth of the theory is contingent on the veridicality of the theoretical components in the theoretical environment, not the degree of surprisingness of the predictions.

A final concern that one may have focuses on the idea that the theoretical work of a particular scientist in a particular case is somehow diminished by the overall theoretical environment (and the successes and failures of others). I have argued that the multiple interdependent attempts that scientists make undermine the idea that novel success is best explained by true theoretical constituents. But why, exactly, should the failures of others impact whatever happens in the case of any particular scientist and their novel success? Could it not be that in some isolated cases novel predictions *are* a reliable indicator of truth?

At this point, we need to recall the perspective from which the issue is evaluated. The predictivists and novelty-based realists have picked a criterion – surprising novel predictive success – that is supposed to provide dependable evidence about true theoretical constituents in science. I have argued against this *criterion* rather than the efforts of any particular theorist in a particular case: the justification of the criterion is undermined by the multiple interdependent attempts that scientists make at theorizing. This does not entail that the efforts of any particular scientist who pursues novel success must be doomed. Rather, the argument establishes simply that surprising novel success is not a criterion that reliably distinguishes the true from the false in scientific practice. There may be other epistemic criteria that we can use to better recognize truth irrespective of the novel prediction versus accommodation distinction, and properly adjudicate between cases where

[77] This is not to take a stance against Bayesian confirmation theory, where the use of evidential probabilities in evaluating theories is standard practice, but rather contribute more to a case in favor of how the position is usually conceived. The Bayesians typically set their probabilities based on the theory, evidence, and background knowledge as such, denying that novel predictions have special value (see Howson & Urbach 1996; Steele & Werndl 2013). Our results provide support for the view that novelty should not be included in confirmatory calculations, as many Bayesians argue.

realist commitment is appropriate and where it should be withheld, whether or not novel predictions were made (see Chapter 6 for further discussion).

5.4 CONCLUSION: VERDICT ON NOVELTY-BASED PREDICTIVISM

With this, we are ready to conclude our investigation of the negative and the positive predictivist theses. It is time to draw overall conclusions:

First, we have found no compelling evidence for the claim that novel prediction provides an important, general advantage over accommodation in science. The negative predictivist thesis claims that novel prediction is better because accommodation is associated with negative epistemic consequences (which arise from overfitting, fudging, and hypothesis hunting). However, we found that novel prediction itself is associated with multiple problems, including p-hacking and publication bias, which do not necessarily leave either novel prediction or accommodation significantly better or worse than the other in the current context. The positive predictivist thesis claims that novel prediction has an explanatory advantage over accommodation: only novel prediction calls for a truth- or reliability-based explanation. However, we found that given the multiple interdependent attempts that scientists make at theorizing, there is no need for a truth- or reliability-based explanation in cases of novel success. There is thus no explanatory advantage to novel prediction over accommodation either, where only one but not the other calls for a realist explanation.

Second, we did find evidence that novel prediction provides some advantage over accommodation in certain cases where we are prevented access to evaluate relevant features of the theory or the evidence as such (e.g. perhaps certain cases of statistical model selection and cases of epistemic opacity). In scientific practice, however, these are cases where our epistemic situation is deeply uncertain, and novel prediction itself does not do much to alleviate that uncertainty. Novel predictive success may sometimes provide modest inductive support that we could be on the right track, but given that we are not in a position to evaluate *why* the theory or the model at hand is successful, we are left with a substantial degree of uncertainty. Moreover, the solution to that uncertainty can often simply be more accommodation (i.e. the use of more data).

Third, despite the negative conclusion on novelty-based predictivism, it is fair to point out that we also did not find much support for the claims of anti-predictivists such as Dellsén (forthcoming) and Harker (2008). Dellsén argues that accommodation may be superior because novelly predicted evidence is more uncertain due to concerns about fraud. However, the study of statistical hypothesis hunting cases shows that accommodated evidence is often uncertain as well, and scientific fraud is generally thought to be relatively rare (see Chapter 4). Harker (2008) argues that underlying our

preference for novel prediction is a preference for progress and simplicity. We have seen, in contrast, that what appears to count for more in science is simply more evidence rather than progress as such. New theories that progress beyond old ones may very well still be predominantly wrong, as the case of Hess illustrates.

We have been left with multiple outstanding problems. 1) The intuition that prediction is an important epistemic indicator in science remains a powerful one (cf. Chapter 1). Is it really the case that there is nothing to this intuition? Is there no epistemic role for prediction in science? 2) In each of the cases we have discussed, scientists have apparently been able to move beyond a state of uncertainty, and select theories that do, according to them, warrant more confidence (e.g. replicable statistical findings, plate tectonics). Based on what do they do this if not use-novel prediction? What are important epistemic indicators in science? 3) Based on our investigation, the novelty-based defense of scientific realism fails. This appears to leave the anti-realist in a considerably stronger position. Is this the correct result about the epistemic status of current scientific theories? 4) We have rejected the idea that contingent novel prediction is epistemically superior to accommodation. However, there is also an entirely different approach to the prediction versus accommodation issue: the logical approach (see Chapter 2). This approach holds that there are certain issues that limit the confirmatory force of logical accommodation, and offers logical prediction as the solution. Could the logical approach offer solutions to our outstanding problems?

6 THE PREDICTIVE VIRTUES APPROACH

Novel predictive success was revealed an undependable epistemic indicator in scientific practice. However, there is one approach still standing in the prediction versus accommodation debate: the logical approach. This chapter explores and further develops this approach.

Worrall (2006) introduced the logical approach to predictivism into the philosophical literature. Worrall agrees with the novelty-based predictivist that there is an important epistemic role for prediction in science. However, he argues that prediction should be understood in a logical rather than contingent sense (i.e. in terms of the empirical results that the theory entails or fits).[78] Logical prediction is distinguished from logical accommodation, which for Worrall is about parameter-fixing: a theory logically accommodates evidence e when it leaves open a free parameter that needs to be fixed based on e. A theory logically predicts evidence when it has empirical consequences that do not depend on this type of 'built-in' relationship between the theory and the evidence.

At the heart of Worrall's logical theory is the idea that the source of scientific predictions are scientific theories (and models). In cases where predictive successes arise, an advocate of the logical approach to predictivism argues that credit should be given where it is due: to *the theory* that had the capacity to generate the successful predictions. Scientists themselves, in so far as they achieve predictive success, achieve it based on the theories and models they have built. In evaluating scientific predictions and their epistemic implications, we should thus focus on scientific theories and their predictive performance rather than the contingent issue of whether a particular scientist used some evidence in the construction of the theory. This chapter examines and develops this idea as a solution to the problem of the epistemic role of prediction in science.

The chapter proceeds as follows. In section 6.1, I re-evaluate Worrall's theory in light of the evidence that we have uncovered in the course of this study. I argue that the theory suffers from multiple problems that indicate that it is inadequate as such. However, the underlying idea that the predictive performance, or *predictivity*, of the theory is what counts from the epistemic point of view is worthy of further development. In section 6.2, I introduce a new framework that develops this view: *the predictive virtues approach*. I argue that this approach is able to overcome the difficulties of Worrall's account and thus represents a better way forward in developing the

[78] Worrall (2006, 2009, 2014) focuses on cases where the theory deductively entails the evidence (with appropriate auxiliaries). Others have observed that Worrall's theory can be interpreted to also concern cases where the theory 'fits' evidence rather than strictly entails it (e.g. statistical cases) (see Mayo 1996, p. 259; Schurz 2014). In the context of this study, the logical approach is interpreted in this broader fashion (see also Chapter 2).

logical approach as a solution to the issues of predictivism. Section 6.3 discusses whether Worrall's criterion of independent testability should be included among the relevant predictive virtues of scientific theories. I argue that it is not needed: the predictive virtues approach captures what is important about this criterion in a more satisfying way. Section 6.4 addresses the question of whether predictive virtues can replace novel prediction in the defense of scientific realism. Sections 6.5 and 6.6 add further clarifications in relation to adjacent debates in the philosophy of science, concerning explanation and scientific progress. Section 6.7 responds to outstanding issues and potential objections, and section 6.8 concludes the chapter.

6.1 WORRALL'S PREDICTIVISM REVISITED

Worrall (2006, 2009, 2014) defends an innovative confirmation theory that is based on multiple original insights about the prediction vs. accommodation distinction. Worrall distinguishes between two types of confirmation: *conditional* and *unconditional*. A theory gains conditional confirmation in cases where it makes logical accommodations, i.e. when the free parameters of the theory are fixed based on the evidence. The confirmation in these cases is limited: it applies only to the particular version of the theory that depends on the fixed parameters. For unconditional confirmation, the theory requires logical predictions, i.e. *independently testable* or *natural* consequences. In exploring the logical approach to predictivism, we must first (re-)evaluate Worrall's theory as the forerunner (and frontrunner) in the field. Does Worrall's theory provide a successful account of scientific confirmation and the issue of predictivism? Furthermore, could Worrall's account be of help to the scientific realist?

With regard to both of these questions, I believe that we must answer negatively, based on the evidence we have uncovered in the course of this study. There are three problems with Worrall's theory that indicate that it does not adequately represent how scientific confirmation works (or should work):

First, despite explicitly rejecting the inclusion of contingent factors in confirmation theory, Worrall's theory ends up inheriting the issues of contingent novel prediction that were discussed in the previous chapter. This is a consequence of the specific way in which Worrall defines logical prediction, juxtaposing it with logical accommodation: a theory accommodates evidence when it leaves open free parameters and it predicts evidence when it entails evidence that is not needed for parameter-fixing (see Worrall 2014, p. 55). The problem is that any use-novel (as well as temporally novel) evidence will not have been used for parameter-fixing, which meets Worrall's criterion for logical prediction and unconditional theory confirmation. Yet, we have seen that use-novelty is not a dependable epistemic indicator in science: the relationship between use-novelty and truth is deeply

contingent in science (see Chapters 4 and 5). Reliance on Worrall's logical criterion of empirical success would thus have us increase confidence in a large group of theories for which that does not appear epistemically warranted. From the point of view of confirmation theory, this is not an optimal result, and it leaves more to be desired.

Second, Worrall's theory contains one notion that, in some cases, might be leveraged to answer to the first problem: 'naturalness.' Worrall could argue that not all (use-)novelly successful theories are equally natural, and hence that we should refrain from committing to certain novelly successful theories.[79] However, this suggestion reveals another problem in Worrall's account: the notion of naturalness is vague, and perhaps hopelessly so. We are given no criteria or indicators for how to compare the degree of naturalness of different theories and theoretical frameworks. Worrall acknowledges this limitation. He writes: "Like all those philosophers of science (pretty well all of them!) who have invoked naturalness or its close relatives 'unity' and 'simplicity' I am unable to provide an adequate, non-circular, characterisation. All I can do is try to elicit the intuition by pointing to examples." (see Worrall 2014, p. 57, fn. 5) This is rather unsatisfying, both from the point of view of confirmation theory as well as in terms of employing Worrall's theory in the defense of scientific realism. One benefit of novel prediction is that it is a criterion that often applies relatively uncontroversially, as is evident in the myriad of case studies in the philosophy of science that identify novel predictions in scientific practice. The notion of naturalness appears much more difficult to operationalize, threatening to leave confirmation theorists (and scientific realists) on even murkier grounds than with contingent novel prediction.

Finally, Chapter 2 introduced further issues which indicate that Worrall's theory may not be successful in capturing the logic of scientific confirmation. Two problems stand out. A) As argued convincingly by Schurz (2014), Worrall's distinction between conditional and unconditional confirmation appears flawed: even if a theory contains a free parameter, it may constrain the results that the parameter can be fixed to in many ways. In these cases, the fact that the empirical results fit these constraints should count at least for some overall confirmation for relevant constituents of the theory. Only in cases where the parameter can fit *any* empirical result does it become inert from the point of view of confirmation. The idea that there are two different types of confirmation thus appears redundant: we could get by with just one type of confirmation that measures the confirmatory impact of the evidence on the theory, as suggested by Schurz (2014).[80] B) There are cases in

[79] This may be easier in some cases than others. For example, Hess's mantle-based theory of continental drift does not appear any more unnatural than various other theories (either successful or unsuccessful) in the continental drift controversy. Similarly, novelly successful but ultimately spurious statistical effects may not appear unnatural at all despite their falsehood.

[80] Schurz (2014) does further distinguish between 'partial genuine confirmation' and 'full genuine confirmation,' but the former is a limited case of the latter in his account, rather than a distinct type of confirmation altogether.

science where fixing theories based on the evidence appears perfectly acceptable, and should be seen as counting for support for the overall theory in question. Consider again the case of the team of epidemiologists who observe a strong correlation between estrogen levels and a response to a particular novel treatment (see section 2.4). In such a case, using the evidence to fix a theory according to which the treatment interacts with estrogen to provide a cure to the particular disease appears epistemically justified. (For further examples, see Mayo [1996, 2014]; Steele & Werndl [2016, 2018].)

These difficulties indicate that Worrall's theory is yet flawed. We may continue to observe that some of these difficulties appear to challenge contemporary theories on the prediction versus accommodation issue also more generally. A chief problem with current theories that has been revealed by our investigation is their inability to distinguish appropriately when more or less confidence is warranted in scientific theories. Worrall's theory suffers from this problem, but it also threatens its chief rival, Mayo's (1996) severe testing account. We saw that the severity account appears to deliver flawed results in the case of statistical hypothesis tests (see Chapter 4). But, in a more general sense, similarly to Worrall's account, the severity account does not contain much resources to establish criteria for comparing which test are more or less severe outside the statistical context; this is left to a significant extent on an intuitive level. For example, it is not clear what the severity account would say about the testing of the Vine-Matthews hypothesis and its impact on the confirmation of Hess's mantle convection theory. Intuitively, the test appears highly severe in probing the possibility that Hess's convection-based theory is in error, but on the other hand the test was not considered to provide strong confirmation for Hess's theory within the scientific community. Absent more precise criteria for how to measure test severity, the account threatens to become trivial: we should believe theories that are unlikely to be false (where what makes theories unlikely to be false is determined by some further unspecified criterion of what makes particular tests severe in a particular context). In the case of Mayo's theory as well, what appears to be missing are more precise epistemic criteria that enable us to make appropriate distinctions between cases where more or less confirmation is warranted.

In the remainder of this chapter, I propose a way forward, addressing these issues. I will build on the logical approach to confirmation and predictivism.

6.2 INTRODUCING PREDICTIVE VIRTUES

In the course of this study, we have evaluated multiple perspectives on the role of predictions in confirming scientific theories. We have been left with multiple puzzles. On the one hand, predictions of different types (both logical and contingent) are ubiquitous in science, and the intuition that they have something to do with the confirmation of scientific theories is powerful.

Scientists constantly develop new theories and make observations and experiments to test the empirical consequences of their theories. The proposition that this is in some sense important for scientific confirmation is impossible to deny. On the other hand, the proposals in the philosophy of science to circumscribe the appropriate kind of predictions that count for scientific confirmation have been unsatisfying. We have seen that the criteria that philosophers have proposed to evaluate scientific predictions do not carve epistemic reality in any reliable way.

In what follows, I suggest a way forward. First, a few points of agreement with the previous proposals. The advocates of the logical approach to predictivism are surely correct in that prediction in the logical sense matters in science. Scientists look for theories that 'predict' successfully, i.e. theories that have (with appropriate auxiliaries) empirical consequences that match observations and experiments (see section 2.4). The advocates of the logical approach are also correct in that contingent factors should *not* be referenced in a theory about scientific prediction and confirmation, at least not in any prominent role; to do so is to succumb to the contingency problem raised in the previous chapter. A final point that I agree with in the previous proposals is that there are *different kinds* of predictive success in science. The previous accounts highlighted use-novel, temporally novel, independently testable, and natural predictive successes. However, the problem with these notions is that they do not appear to be getting at the heart of what it means for *a theory (or model)* to be predictive successful. Temporal novelty and use-novelty are incidental to the theory and its relationship with the evidence. 'Naturalness' as a characterization is vague, and it is also clearly not a necessary precondition for predictive success (consider, for example, a climate model that is highly complex but is nonetheless useful for predictive purposes). Arguably, only the generation of independently testable consequences speaks to some type of predictive virtue that could be attributed to the theory itself. Yet, independent testability too is still a very loose criterion on its own, and it is insufficient to demarcate properly between cases where more or less confirmation is warranted in science (see Chapters 4 and 5).

I suggest that instead of the loose and incidental measures of predictive success, we should turn our focus directly on the predictive performance of the scientific theory itself. The purpose of science in general is to generate theories and hypotheses about what the world is like. These theories and hypotheses generally extend beyond what is directly observable (e.g. they make inductive generalizations and posit unobservable laws, mechanisms, and entities), so they are tested indirectly by checking if they fit empirical results in the world. If empirical results bear out what the theory indicates should be the case, this provides confirmation to the theory. If the empirical results fail to show what is expected based on the theory, this provides disconfirmation to the theory. What matters for scientific confirmation, I suggest, is that theories are predictive in this sense, i.e. they are able to fit empirical results in a way that the relationship between the

theory and the evidence is appropriately 'predictive' (specified further below).[81] From the point of view of philosophical theories of scientific confirmation, what are needed are criteria that can be used to evaluate how well theories perform predictively in the logical sense.

Logical confirmation theories have sometimes been thought to lack the resources to do something like this, i.e. they have been thought to lack the resources to distinguish appropriately which logical consequences of theories provide stronger confirmation (see, for example, Alai 2014; Schurz 2014). However, this is not an unavoidable feature of logical confirmation theories: it is possible to make logical confirmation theories more discriminatory by appealing to predictivist ideas (e.g. Worrall 2006, 2014; Alai 2014; Schurz 2014). I propose that we can do this most effectively if we specifically set the scientific theory as the object or unit whose logical predictions are evaluated for theory confirmation. Recognizing the theory, and its performance with empirical results, as what fundamentally counts, we can identify a number of logical prediction criteria that impact whether more or less confirmation is warranted for the theory. These criteria ultimately determine the degree to which a theory is confirmed by the evidence, rather than the contingent accident of which evidence was or was not used by the theorist.

To this effect, I want to introduce four predictive 'dimensions' or 'virtues' more closely. These virtues probe in different ways the capacity of the scientific theory itself to make logical predictions about the world, capturing distinct senses in which we may consider that a scientific theory is more or less predictive. These four virtues are, to a certain extent, already implicit in the findings that we have uncovered in the course of this study, but this has not been made obvious yet. Some of the virtues have also recently been highlighted in some form or another by novelty-based predictivists (see Alai 2014; Vickers 2019). However, I believe that the problem with these theories is that they have focused on the "wrong" kinds of predictive success, and understood scientific prediction altogether too narrowly. Previous predictivist theories either invoke criteria that are not relevant (e.g. surprising novel success) or miss the underlying logical sense in which predictions matter for confirmation. I suggest that a more detailed understanding of what it means for a scientific theory to be predictively successful in the logical sense enables us to achieve a closer match between philosophical theories of scientific confirmation and the epistemic reality of scientific practice. (As we now move to focus on logical

[81] Schlesinger (1987) has approached the issue from a similar angle. Schlesinger argues that "predictive power, or the capacity to serve as a means for prediction" (ibid. p. 33) is what matters for determining the degree of support for the theory. He defends the following epistemic principle: "[t]he hypothesis warranted by the first few results as the one to be adopted, keeps being the hypothesis whose adoption is required by increasingly larger sets of results" (ibid. p. 35). Instead of a single principle or rule, I propose that we think of the predictivity of a scientific theory in terms of multiple dimensions or virtues. Perhaps the closest precedent to my approach is Fahrbach (2011a; 2011b; 2017), who argues that the amount, diversity, and precision of scientific evidence has increased greatly in recent decades, making current theories more predictively successful than past theories.

prediction in particular, i.e. we consider predictions as the empirical results that the theory fits or implies, I will generally omit the term 'logical.' 'Prediction' in what follows always means 'logical prediction' unless otherwise stated.)

a) *Veridicality (/accuracy)*. First, we should set a very basic criterion: true predictions provide confirmation to the theory while false predictions provide disconfirmation. From the point of view of logical theories of confirmation, theory confirmation depends on the empirical results that the theory fits (i.e. empirical observations and outcomes of experiments). If the theory fits or implies some true empirical result e_x, i.e. the theory logically predicts e_x so that e_x follows from the theory and background assumptions, e_x counts for some degree of confirmation for the theory. If the theory fits or implies a false empirical result e_y, e_y counts for disconfirmation. In determining the degree of confirmation for the theory, we need to at least take into account the veridicality of the empirical predictions it makes.[82]

The relevance of considering the veridicality of predictions from the point of view of the predictivity of the theory itself can be illustrated further with the example of false but novelly successful theories (e.g. Ptolemaic astronomy, the zymotic theory of disease). If we examine these cases from the point of view of the predictivity of the theory, I suggest that an important problem that is revealed in these theories is that they also make a number of observable *false* predictions, which has also been highlighted recently by Vickers (2019).[83] A false theory such as Ptolemaic astronomy may fit some true empirical results that end up as surprising novel predictive successes. However, from the logical point of view, we can see that the theory also makes a number of important false predictions, including the very wrong prediction that the planets of the solar system go through retrograde motions. In evaluating the degree of confirmation for a theory such as Ptolemaic astronomy, I propose that we need to take into account the veridicality of its empirical predictions rather than simply its surprising novel successes.[84]

b) *Specificity*. One of the chief attractions of predictivism has been that it has enabled the predictivists to move beyond the basic dimension of veridicality. All confirmation theories hold that, other things being equal, it is better to make true rather than false predictions. However, not all predictions

[82] Vickers (2013, p. 196) also identifies a further way to evaluate predictive accuracy, referring to "the degree to which the prediction matches experiment(s)." That is to say, the dimension of veridicality or accuracy can be further specified to take into account not only whether the predictions of the theory are true but how closely they approximate empirical results.

[83] Vickers (2019), however, continues to demand surprising novel success in this context, which I have rejected.

[84] A well-known issue in logical confirmation theories concerns how we distinguish between relevant and irrelevant logical consequences of the theory. Schurz (1991; 2014, p. 93; see also Schurz and Weingartner 2010) has developed a solution to this problem that appears attractive to me. Schurz's account ties relevant empirical results to the content elements of the theory, so that for example content elements whose content could be any sentence whatsoever are excluded. The predictive virtues approach also introduces a number of further discriminating criteria for logical predictions and predictivity, which are discussed below (see also sections 6.3 and 6.7).

are made the same. For example, if we go back to Velikovsky's prediction that Venus is "hot" rather than "cold" (see section 3.2), it appears that this prediction is not very powerful. Even if the theory at hand is false, we can see that it is not very difficult to get a prediction like this correct. It seems, intuitively, that the more impressive the prediction, the more it should confirm the theory.

The surprisingness criterion in predictivism was meant for this purpose: surprising novel predictions confirm the theory more strongly than unsurprising predictions. Unfortunately, we found that surprisingness is an undependable epistemic criterion in scientific practice, so this advantage appears lost. However, there is another, closely related notion that could be better. I suggest that the predictivist's surprisingness criterion can be replaced by another criterion that applies to the predictions of the theory itself: *specificity* (of testable and observable predictions) (see also Alai 2014). Predictions differ in the degree to which they are vague versus specific, i.e. in the degree to which they set constraints on the world. The underlying problem with Velikovsky's prediction is its *un*specificity. This prediction only puts a vague constraint on the world, holding that the world must be one way rather than another from two logically exhaustive options. More specific predictions constrain the world more strongly, demanding more from the theoretical mechanisms from which they are derived.

The confirmatory relevance of the logical specificity of predictions has been defended recently by Alai (2014), who proposes the notion of 'a priori improbability' to measure logical specificity. Alai argues that the a priori improbability of a prediction matters for confirmation because we can see that the class of theories that fit unspecific results about some aspect of the world is much larger than the class of theories that fit specific results. In the former type of case, there are countless numbers of false theories that fit the unspecific result, so making such a prediction does not alleviate the concern that we are only dealing with one of these false theories. In contrast, more specific predictions do alleviate this concern, calling for a more substantive explanation in terms of the theory latching on to reality.

To measure the specificity or a priori improbability of a prediction, Alai suggests that we refer to the class of possibilities it excludes: the more possibilities are ruled out, the more specific and confirmatory the prediction.[85] He notes that comparisons between degrees of specificity may also oftentimes be clear enough in scientific practice without the need to introduce a specific quantitative measure. For example, we might have a theory that predicts the existence of a) a planet in the Universe, b) a planet in

[85] Alai (2014, p. 308) proposes that in dealing with a space of finite possibilities (e.g. basic vocabulary of a language, minimal differences detectable by instruments), a priori improbability can be measured by counting the excluded possibilities and in the case of an infinite space by using the class inclusion relation. The measure is thus necessarily tied to a language or a conceptual scheme. Alai argues, however, that the results may often be fairly definite given other accepted theories that constrain the appropriate space of possibilities in a given scientific context.

a particular solar system, or c) a planet with its exact mass and orbit in a particular solar system (see Alai 2014, p. 307). We can see that the a priori improbability of the prediction increases as we move from a to b to c.

For the purposes of what follows, we can take predictive specificity from the more general point of view, where we tie predictive specificity to the appropriate scientific context. Predictions can be seen as more or less specific depending on background assumptions in the context. In any scientific context, there are many observable results that can be recognized as trivially unspecific and can be established without the help of a scientific theory (e.g. "the continents exist," "the seafloor exists," "the Sun will rise tomorrow"). Predictions become more or less specific against such background. For further theoretical development, Alai's proposal on a priori improbability provides one option for how we could develop a more precise quantitative measurement of predictive specificity. However, in contrast to Alai, I believe that predictive specificity counts only if it is achieved by a theory that is properly seen as predictively virtuous, which produces some significant differences between our two accounts (see section 6.3 for further discussion).[86]

c) *Scope*. Demanding that scientific theories make specific, veridical predictions enables us to pick out more or less confirmatory single predictions while keeping within the logical framework. However, there is a sense in which this still covers only a very narrow range of what makes scientific theories predictive, and what appears to count for overall theory confirmation. Hints of another type of predictive virtue or dimension were seen in Chapter 4. We found that confidence in scientific theories also grows *the more evidence* there is for them; e.g. the better they hold in replication attempts, the more data there is to back them up, or simply the more true empirical results they fit in their domain. This also appears to have something important to do with both confirmation and logical prediction. For example, in something like a replication attempt for a hypothesis about a population level effect, the scientific theory as such entails that the replication attempt should show certain empirical results if the theory is to be empirically supported. The theory predicts that the hypothesized effect should show up also in the replication attempt (and in any further such attempts). If these logical predictions of the theory fail, the theory is disconfirmed in consequence; if the predictions are successful, the theory becomes more confirmed. By being predictive in this sense, i.e. by setting themselves up for a range of empirical tests, scientific theories can display stronger predictive performance with regard to the world.

To capture this type of predictive virtue, I suggest that we add the notion of *scope*, which is relativized to the domain to which the theory is applied. All scientific theories have a particular target or domain that they are developed for. The more (correct) empirical results the theory fits within that

[86] Instead of predictive virtues, Alai (2014) demands 'inessentiality' and 'heterogeneity' from the a priori improbable prediction, but I suggest this leads to the wrong epistemic verdict in important scientific cases (see section 6.3).

domain, the more predictive it can be seen with regard to that domain, and thus it can ultimately become more confirmed. For example, consider a linear regression model that attempts to capture a particular data-generation mechanism in the world. The more data points or datasets that the model fits, the more predictive it can be seen in relation to the underlying data-generation mechanism (see also section 6.7 for further discussion on the relevance of model selection methods to the evaluation of predictivity). Or, for another example, consider a theoretical mechanism that attempts to account for geological phenomena related to the oceans and the continents (e.g. Hess's theory of continental drift). The more correct empirical results the theory fits in this domain, the more predictive it is with regard to this domain.[87] In considering the predictivity of a scientific theory, we should demand that it demonstrates scope in addition to veridicality and specificity.[88]

d) *Counterfactual depth*. Veridicality, specificity, and scope enable us to begin to evaluate the predictions of scientific theories from a more comprehensive perspective. However, there is still a sense in which we have not captured all of what makes scientific theories predictive with regard to the world. We are still focused on isolated predictions, whether there be one or more of them. I believe that this has been a general problem in the predictivism debate, which has had a very narrow focus on singular predictions (e.g. surprising novel successes). If we consider a good scientific theory, however, such as the general theory of relativity, there is a sense in which the focus on these one-off surprises or successes greatly undervalues the capacities of this theory. This theory is capable of predicting much more than just isolated results. Rather than simply having the ability to predict isolated results, what is impressive about this theory is that it is able to act as *a continuing source* of predictions about the world. Using the general theory of relativity, we can continually predict how changes in the mass of any large object (e.g. a planet or a star) impact the degree to which it bends light, produces time dilation, precesses in orbit around another large object, etc. In other words, this theory does not just imply a collection of individual results, but its theoretical postulates enable us to predict what would happen in different kind of contexts and circumstances (and, crucially, such predictions have turned out correct as far as they have been tested). This, I propose, is one particularly important

[87] Determining the appropriate scope that should be demanded from a particular theory is partly a matter of background knowledge and partly a consequence of the kind of theory that one advocates. For example, if one proposes a theory of continental drift that is based on a mechanism of mantle convection, the domain of the theory is set naturally by the theoretical mechanism that one postulates to exist in the world. The theory holds that there is mantle convection, which as such has multiple consequences with regard to the oceans and the continents (with further accepted auxiliaries). With the domain of the theory set, it becomes possible to test the theory to various degrees of *comprehensiveness* within its domain.

[88] Scope is related to but distinct from the explanatory virtue of *unification*. A theory that has great unificatory power will have broad predictive scope, but predictive scope can also increase without unification. For example, a machine learning algorithm may produce a highly disunified model which nonetheless has broad predictive scope. I suggest that making this distinction could motivate interesting comparisons between the epistemic value of prediction and explanation (see section 6.5.)

property or feature that demonstrates an underlying predictive virtue in this theory that we need to take into account in our approach.

To capture this sense of predictivity, I suggest the notion of *counterfactual depth*, which refers to the degree to which the theory can be used to make predictions about empirical results based on counterfactual inferences. Counterfactual inferences provide answers to *what-if-things-had-been-different* questions, i.e. questions about how a target system would behave or would have behaved under various interventions or changes (see Woodward 2003). When theories expand into the counterfactual dimension, their predictivity expands in a powerful way. By enabling us to predict the results of specific changes and interventions, scientific theories become a potentially endless source of predictions about the world. In terms of the counterfactual depth of the theory, we should at least demand that the theory is 'invariant' (in the sense of Woodward 2003, Chapter 6); i.e. that it does not constitute an accidental generalization.[89] Scientific theories can, however, progress well beyond this, which can make them more and more impressive in terms of their predictive performance.[90]

Taking these four predictive virtues or dimensions, we can set more discriminating criteria for logical prediction and predictivity. In order for a theory to be in any sense predictive in relation to the world, it must at least fit correctly some empirical results. The empirical results that the theory fits must be at least be specific enough so that something is ruled out about the world. (A theory that does not entail or imply anything, or alternatively implies anything, does not predict; it is unfalsifiable.) To have scope, the theory must also go beyond the evidence as such; i.e. it must contain some theoretical postulates that state more than just the empirical results as such. Finally, the theory must at least be counterfactually invariant (at least in so far as the theory does not constitute a simple empirical generalization). Given the close connections between the virtues, in that they appear to capture different *aspects* of what it takes for a theory to be predictive, we might also sometimes want to refer simply to *the predictive power* of the theory rather than predictive virtues individually, analogous as to how we might talk about both

[89] A paradigm example of a relationship that is not invariant (and hence not appropriately seen as predictive) is the relationship between a falling barometer level and storms. An intervention on the barometer does not impact storm occurrence, so a theory that argues that falling barometer levels cause storms is non-predictive (in the sense of logical predictivity). For further discussion on how to make the notion of invariance more precise, see Woodward (2003, Chapter 6).

[90] Similarly to scope, counterfactual depth contributes to confirmation through increased testability: the ability to accurately predict the results of specific interventions and changes enables the theory to be tested in a very thorough way. Counterfactual depth appears to contribute to confirmation also similarly as specificity, in that theories that achieve counterfactual depth are much rarer than theories that lack counterfactual depth. Finally, counterfactual depth also provides evidence about causality and causal mechanisms. Counterfactual depth appears in some sense to demonstrate a particularly strong *combination* of both specificity and scope. However, the virtue is also clearly distinct in that specificity and scope can grow without counterfactual depth (e.g. with empirical generalizations). This perhaps reveals some degree of hierarchy in the predictive virtues we have introduced. Further studies are needed to explore more closely the relationships between the virtues.

'explanatory power' and 'explanatory virtues.' However, for the purposes of what follows, I propose that we use the more nuanced notion of predictive virtues, as this allows us to make more precise distinctions between different kinds of predictive success in scientific practice.

My proposal in accounting for the relevant way in which prediction matters for scientific confirmation is that predictive virtues, and the underlying sense of logical predictivity of the theory, is what counts rather than novel predictive success. Scientific theories are evaluated based on the empirical contact they make with the world, i.e. based on their logical predictions. What we want in science are theories that are good predictors of empirical results, theories whose capacity for prediction is probed in multiple ways. Generating theories that are able to achieve strong predictive performance is not easy, as we have seen in the course of this study. Thus, evaluating whether theories are genuinely predictive cannot rest on singular, contingent successes. The evaluation of the predictions (and the underlying predictivity) of the theory must proceed in a more comprehensive manner. To assess whether a particular theory is a good predictor or not, we need to consider its multiple predictive virtues. Does the theory make true predictions, or are there relevant false predictions? Are the predictions of the theory specific or vague? What is the domain to which the theory is applied, and to what extent does the theory predict empirical results within that domain? Does the theory expand into the counterfactual dimension? Theories that perform well in light of these criteria become, other things being equal, more confirmed than theories that do not.[91]

To specify, I emphasize that my purpose in introducing these four predictive dimensions or virtues is not to claim that these necessarily capture *all* aspects of scientific prediction, nor do I wish to commit to such a position. It is possible to think of more properties of the theory that reveal different types of predictive virtues in scientific practice. For example, one such property is the computational cost involved in deriving predictive consequences from a theory or a model.[92] A model that is not perfectly accurate but whose predictions are not too costly to compute could very well be preferable in certain contexts to an alternative that is more accurate but too

[91] In the account that I have proposed, confirmation thus depends on particular predictions and ultimately the overall predictivity of the theory from which they are derived. One possible way to represent the distinction between the overall theory and particular predictions is in terms of Bayesianism. The Bayesian *prior* could be set based on an assessment of predictivity, and *conditionalization* is applied with individual predictions (see also section 6.7). The degree of confirmation provided by an individual prediction can vary based on three factors: accuracy, specificity, and counterfactual depth. The theory becomes more confirmed in virtue of more such instances of confirmation being added in the context of the appropriate domain. While I believe that this covers much of what makes single predictions more or less confirming, I do not wish to commit to the position that these must always be the only factors that count in single instances. For example, we also need to take into account in a more specific way which specific constituents of the theory hinge on the accuracy of a particular prediction (see Schurz 2014). Further work is needed to determine even more closely the degree to which a particular prediction confirms particular constituents of the theory.

[92] I would like to thank Petri Ylikoski for making this point.

costly to employ in practice. Further investigations are needed to explore the various ways in which scientific theories can be more or less predictively virtuous, including a comparison of the *epistemic* and *pragmatic* predictive virtues of scientific theories. At this point, what I want to do in particular is to introduce *a framework* in which we begin to consider scientific theories from the point of view of their predictive virtues. The particular set of virtues that I have highlighted appears relevant to addressing the *epistemic* problems that have occupied us in this study (i.e. the problems of predictivism and scientific realism), so I will focus on them here.

In other respects, relevant dimensions of the discussion should remain the same. Background knowledge still counts, as per usual. Theories and their predictions are evaluated in part based on how they relate to previously accepted theories and scientific background principles. In the predictive virtues approach, background knowledge plays a role in determining the context in which the predictive virtues of the theory are evaluated. In this, the approach is similar to other accounts, such as that of Psillos (1999), Worrall (2009), or Mayo (2014), each of whom requires additional background conditions for the application of their respective confirmation criteria. For example, Psillos (1999, p. 102) demands 'maturity' from the scientific context, which is "characterised by the presence of a body of well-entrenched background beliefs about the domain of inquiry which, in effect, delineate the boundaries of that domain, inform theoretical research and constrain the proposal of theories and hypotheses." The predictive virtues recommends in similar fashion that predictions are evaluated within the appropriate scientific context, where certain auxiliary assumptions are appropriately taken for granted (e.g. the observable features of the world, previously confirmed scientific theories, general background principles such as rules about inductive inference).

Another dimension that I propose should remain the same is the selectivist idea that scientific theories are broken up into their constituent parts (e.g. Psillos 1999; Vickers 2013). Scientific theories often contain multiple theoretical constituents, where not all constituents are necessarily contributing equally (and some may not contribute at all). This idea seems to me to be important whether we consider logical or contingent predictions: theories may have inflated theoretical structures that do not play a part in deriving relevant empirical consequences from the theory. Predictions can confirm only those theoretical constituents that actually contribute to the derivation of the predictions.[93]

[93] One example of an irrelevant constituent is Newton's postulate that the center of the Universe is at absolute rest (see Psillos 1999, p. 104). This postulate does not play any part in the testable predictions of Newton's theory, so it is not confirmed by the predictive successes of Newton's theory. In terms of an appropriate selectivist account, we could accept, at least preliminarily, Vickers's (2013) distinction between DEPs and DIPs, and the minimal logical consequences of DIPs that are needed for the predictions of the theory (see Chapter 3). An important difference to the old is that the predictive virtues approach recommends that more of the theory remain intact in cases where we attempt to eliminate idle posits, as maintaining predictivity requires more parts in the theory than isolated novel successes. This

Finally, we may continue to recognize use-novel success and temporally novel success as distinct kinds of predictive success in scientific practice. However, these are simply not very important types of success from the confirmatory point of view. These notions were introduced into philosophy of science because it was thought that they bring additional benefits to confirmation theory over and above logical confirmation criteria. We have seen that these benefits are not realized in scientific practice: novelty does not have dependable epistemic implications. Thus, we should stick to logical criteria. Even though this requires for us to let go of certain ideas that were thought to be advantageous to confirmation theory (e.g. surprisingness), I suggest that this now enables us to introduce another perspective to scientific prediction that brings more new advantages. As far as the contingent novelty-dimension does count in certain isolated cases, I suggest that this is only because in these contexts novel success enables us to measure something more fundamental: the logical predictivity of the scientific theory (see section 6.7).

In the following sections, I continue to further clarify how the predictive virtues approach is applied in scientific cases, and explore potential benefits of (and possible objections to) the approach. However, we can now briefly return to the problems that we have sought to resolve. We have been looking for a confirmation theory that is able to establish more discriminating criteria of success for scientific theories, without succumbing to the contingency problem of novel prediction. The predictive virtues approach can deliver with these desiderata. First, it respects the maxim that only the logical or structural relationship between the theory and the evidence should count in evaluating the theory. To evaluate the veridicality, specificity, scope, and the counterfactual depth of a scientific theory, we only need to examine the theory and the evidence as such (along with relevant background knowledge). Second, unlike previous predictivist theories, the new approach delivers more discriminating criteria of success for scientific theories. It enables us to identify more or less confirmatory individual predictions, and also ties theory confirmation to the overall predictivity of the theory. I suggest that these criteria can be used to distinguish between good and bad cases where there is more or less confirmation in scientific practice, a point that I shall explore further in sections 6.3 and 6.4. Finally, given that there is no prohibition on the use of evidence, we also have criteria that allow for theories that were constructed directly based on the evidence to nonetheless count as more or less predictive with regard to that evidence (and more or less confirmed by that evidence). The next section explores how the criteria that we have set allow confirmation to be allocated in the appropriate way in these cases.

---

can be advantageous in that we are able to view theories as more closely knitted structures while retaining the ability to discount some idle posits. On the other hand, this also means that in the predictive virtues account, theories are more vulnerable to false predictions as a whole, i.e. more parts of the theory may come into question in the case of false predictions. A closer investigation into selectivity and predictive virtues is needed in the future. See also section 6.3 for further discussion on how different kinds of empirical results are handled by the predictive virtues approach.

6.3 ON INDEPENDENT TESTABILITY AND AD HOCNESS

The predictive virtues approach differs in one aspect rather dramatically from previous predictivist approaches: there is no prohibition on the use of evidence, or a general penalty that arises because evidence has been used in the construction of the theory. Unlike in Worrall's account, there also seems to be no apparent logical analogue of this criterion (i.e. a confirmatory discount due to parameter-fixing). Yet, in the course of this study, we have seen that there are clear examples of cases where evidence is incorporated into a theory in an objectionable way. Consider, for example, the addition of free parameters that are constructed to accommodate any evidence from reality (see Schurz 2014). Similarly, we can see that there are always ways to work empirical results into the theory 'by hand' by making more and more parameter adjustments (see Worrall 2009). Surely, evidence that has been incorporated into a theory in these ways is incapable of providing strong or meaningful confirmation. How should these cases be managed in a predictivist confirmation theory?

Worrall (2014), as we can recall, argues that as far as free parameters are to count for the confirmation of the overall theory, they must generate independently testable consequences with the overall theory. I have not appealed to this criterion, and this is because I believe it is not needed. I suggest that the predictive virtues approach entails the appropriate confirmatory verdict about free parameters – unlike Worrall's account (see Schurz 2014). The criterion that captures the confirmatory impact of free parameters in the predictive virtues approach is *specificity*. The epistemic issue that arises with free parameters, I suggest, is the *un*specificity of the consequences of the theory in these cases. If a theory is supplemented with free parameters that are compatible with any number of empirical results, its predictive specificity in relation to those results is low or nonexistent. Accordingly, the theory receives less or no confirmation from evidence that has been incorporated into it by adding free parameters. However, crucially, unlike Worrall argues, this does not mean that there is a categorical difference between different kinds of confirmation, nor that the overall theory is incapable of receiving any support from parameter-fixing. The underlying factor that matters for confirmation is the degree to which the overall theory sets constraints on the free parameters – as argued convincingly by Schurz (2014). If we have a theory that is be supplemented with a free parameter but it nonetheless constrains the contents of the parameter to a high degree (e.g. it entails that the parameter can only take values between, say, 1 and 2, out of multiple possibilities), it can receive a meaningful degree of confirmation if the measured parameter value does agree with the constraints, because the empirical results that the theory fits are rather specific. In contrast, if a theory is supplemented with a free parameter that is compatible with any empirical result, no confirmation accrues to the theory, because the theory does not

contain postulates that constrain the contents of the parameter to any extent. In either case, what matters is the specificity of the predictions of the theory rather than the fact that a free parameter has been used.

Highlighting this aspect of the new approach enables me to introduce further clarifications that distinguish my position from other logical predictivist approaches. A chief motivation behind Worrall's demand for independent testability has been the intuition that evidence that is *needed* in the construction of the theory must somehow be unable to strongly confirm it (see Worrall 2006, 2009; Schindler 2014; Alai 2014). The basic idea is that if we do need evidence to construct a particular theory (or a version of the theory), the evidence could easily have been incorporated into the theory in some kind of ad hoc way; the fact that we needed the evidence in some sense betrays that we do not have a theory that could have been used to predict that evidence. Theorists such as Alai (2014) consider this as a constitutive requirement for novel predictive success and its confirmatory advantages. Alai argues for a criterion of *inessentiality*, where (functional) novelly predicted data must not be used essentially in constructing the theory (i.e. the theory needs to be plausible independent of the novelly predicted data).

I believe that this intuition is misguided: there is nothing wrong with evidence that is essential or needed in the construction of a theory. In fact, I would argue that exactly the opposite is the case: evidence that is essential is essential precisely because if it (along with other essential pieces of evidence) did not exist, there would be no basis to advocate the theory, i.e. there would be no epistemic support for the theory. Consider a simple example: a theorist uncovers evidence e, and generates a theory T that explains e. Other than e, there is no reason to postulate T – i.e. e is essential to the construction of T. Now, it turns out that the methodological setting in which e was generated contained a flaw, and ~e actually obtains. Theory T, it appears, is disconfirmed in consequence. I suggest that if a negative outcome from a particular experiment or observation is capable of disconfirming a theory, a positive outcome from the experiment or observation must also, other things being equal, confirm the theory to some degree.

This point can be illustrated further through a practical example, e.g. the case of epidemiologists that was discussed in Chapter 2. A team of epidemiologists discover a strong correlation between estrogen levels and response to a particular novel treatment, and use this evidence to construct a theory according to which the treatment interacts with estrogen to provide a cure to a novel disease. Other than the correlational evidence, there is no evidence for or against the estrogen theory. The evidence is thus essential to the construction of the theory; yet, it appears (intuitively) to confirm it to a meaningful degree. I suggest that the predictive virtues approach, unlike the predictivism of either Alai or Worrall, entails the correct verdict about this case. The degree of confirmation here depends on the predictivity of the theory with regard to the empirical results. Confirmation increases the more veridical, specific, wide-ranging, and counterfactually deep the estrogen-

theory is in relation to the evidence. For example, if the theory says that higher levels of estrogen 'increase the likelihood' that the novel treatment works, and this really is the case in a certain dataset, this will count for something in terms of confirmation. The more evidence there is, e.g. the theory is tested with more data or more samples, the more confirmed the theory becomes. If the theory is made counterfactually deeper, for example through the introduction of a precise 'law' that relates estrogen levels to the likelihood of response to the treatment, again the better that is from the confirmatory point of view. At all times, what counts is the predictive performance of the theory with regard to empirical results – not who used or needed to use what evidence and when.

Sanctioning the use of evidence in the construction of the theory may still raise suspicions. Even if the predictive virtues approach can account for cases where evidence is used in an appropriate way, what about the inappropriate use of evidence? Surely, there is such a thing as an *ad hoc* modification to a theory. Consider, for example, the now defunct phlogiston theory of combustion. Advocates of the phlogiston theory famously postulated that substances emit an element called 'phlogiston' when they burn. When experiments revealed that some substances actually *gain* weight while burning, some advocates of the theory proposed a novel solution: phlogiston has *negative* weight (see Partington & McKie 1938). Intuitively, such an adjustment does not appear to confirm the phlogiston theory at all.

The problem of how to define ad hocness has proven notoriously difficult in the philosophy of science. There is no generally accepted solution (see Hunt 2012 for a review). Ad hocness, it seems, may come down to scientists' aesthetic tastes (see Hunt 2012). However, the predictive virtues approach does have some things to say about these type of cases. Cases where there seem to be clear ad hoc adjustments are often cases where the theory appears to be clearly *refuted* by a failed prediction (based on the contents of the theory and background knowledge), and an adjustment is then made to *reduce* the predictivity of the theory. For example, in the phlogiston case, the observation that some substances gain weight while they burn directly contradicts a basic postulate of the phlogiston theory, at least as far as it is agreed that there are no such things as substances with negative weight. (If, in contrast, this was accepted and testable in light of background knowledge, the negative weight adjustment could very well be reasonable. Partington and McKie [1938] discuss different versions of the negative weight assumption, and contemporary responses in the scientific community, in more detail.) Thus, the predictive virtues approach delivers the verdict that the original phlogiston theory is disconfirmed in this instance. On the other hand, if we consider a new version of the phlogiston theory, where phlogiston is given a novel theoretical property which shields it from any testable empirical consequences (e.g. negative weight), the new theory fails on the *specificity* criterion. Accordingly, there is no confirmation for the adjusted version of the theory either. The predictive virtues approach is thus able to deliver intuitively

appropriate verdicts: the original phlogiston theory is refuted, and the ad hoc version fails because it is no longer predictive with regard to empirical results.

There are further issues that have been raised against logical theories of confirmation in general, a more thorough evaluation of which falls beyond the scope of the current study. For example, one such issue is the problem of irrelevant tacking (see, for example, Schippers & Schurz 2020). In these cases, an irrelevant hypothesis is tacked to a theory by conjunction (e.g. theory t becomes theory t ∧ h, which logically entails t and whatever evidence there is for it). Critics argue that an advocate of a logical theory of confirmation is committed to the consequence that any such irrelevant hypothesis is confirmed by the evidence that follows from the original theory. Advocates of logical confirmation theories have proposed multiple solutions to this (and other) logical paradox(es), the evaluation of which is beyond our scope.[94] However, we may observe briefly that the predictive virtues approach can also readily incorporate principles that rule out confirmation in these kind of cases. Minimally, all we need is to apply the selectivist criterion, which holds that particular predictions can confirm only those theoretical constituents that contribute to the derivation of the predictions.[95] Another constraint that the new approach sets for irrelevant tacking is the demand for predictive scope, which is tied to the domain to which the theory is applied. Irrelevant parts do not improve the scope of the theory in this sense, so they are ruled out. Further studies are needed to explore how the predictive virtues approach relates to the various logical paradoxes and the alternative solutions to them.

6.4 PREDICTIVE VIRTUES AND SCIENTIFIC REALISM

One of the most important motivations for introducing the notion of novel prediction into the philosophical literature has been its application to the scientific realism debate: the scientific realist seeks appropriate criteria of success to select theories worthy of realist commitment. We have found that the novelty-based defense of scientific realism is unsuccessful, which appears to give the scientific anti-realist the upper hand. To evaluate further the capabilities of the predictive virtues approach, we should consider whether the approach affects the dynamics of the scientific realism vs. anti-realism debate.

---

[94] For one prominent solution to the tacking problem, see Schippers and Schurz (2020). Another famous problem that has been raised against logical theories of confirmation is the 'raven paradox,' which I will not attempt to tackle here. For current purposes, it suffices that we make the conditional claim that as far as problems such as the raven paradox are not insurmountable, the predictive virtues approach is on the table as a potential contender in improving logical theories of confirmation.

[95] Another attractive option is Schurz's (2014) notion of genuine confirmation, which measures whether the evidence is probabilistically relevant to theoretical constituents. However, a significant difference between Schurz's proposal and mine is that he holds that the confirmation of statistical hypotheses requires a use-novel dataset – a point that I have explicitly denied based on the findings of Chapter 4. So, Schurz's proposal will require some adjustment if it is to be made compatible with the predictive virtues approach.

Could the predictive virtues approach be of help to the scientific realist, and replace the novelty criterion of success?

First, let us return to the challenge that was raised against the novelty-based defense of scientific realism: the criterion of surprisingness is undependable in the contingent context of scientific practice where multiple interdependent attempts at theorizing are made. It recommends belief based on contingent accidents that do not provide dependable evidence about the truth of the theory. Can the predictive virtues approach do better?

Consider, as tests for the new approach, Einstein's theory of general relativity and Hess's mantle-based theory of continental drift. Both of these theories achieved highly surprising and impressive novel predictive successes. Yet, only the former was (and is) considered highly confirmed by the scientific community. What does the predictive virtues approach imply about the confirmatory status of these theories? The predictive virtues approach holds that in order to decide whether and to what degree we should believe either theory, we need to evaluate it in terms of its veridicality, specificity, scope, and counterfactual depth. Take, first, general relativity. In terms of veridicality, this theory is among the greatest in science: so far, it has never failed under experiment. Its scope is also enormous: it generates predictions about any sufficiently large gravitational effects and phenomena in the Universe (with possibly the exception of the center of black holes and the very first stages of the Universe). The specificity of its predictions is also unrivalled in the history of science (apart from quantum mechanism), as demonstrated, for example, by the incredibly precise prediction of the perihelion of Mercury data and later gravitational waves. Finally, the theory supports innumerable counterfactual inferences. For example, as pointed out before, the theory predicts how changes in the mass of an object such as a planet or a star impact the degree to which it bends light, produces time dilation, precesses in orbit around another large object, etc. The predictive virtues approach delivers the appropriate verdict that general relativity is highly confirmed due to its predictive performance with regard to the world.

In contrast, Hess's mantle-based theory of continental drift achieved much less impressive predictive performance. This theory had two important true consequences: seafloor spreading and continental drift. Yet, it also had multiple issues concerning its veridicality, specificity, scope, and counterfactual depth. The theory struggled to demonstrate that mantle convection cells could produce enough stress to drive plate movement (see Forsyth & Uyeda 1975, pp. 166-167). The theory did not account for many other relevant phenomena in its domain, such as places where oceanic ridges migrate and collide with oceanic trenches and triple junctions where three continental plates meet (Forsyth & Uyeda 1975; see also Franklin 2012d). The theory could be applied to derive correction predictions at certain sections of the ocean floor – e.g. oceanic ridges where two continental plates move away from one another – but it did not fit empirical results about other important geological phenomena that fell within its domain. Very soon, a better

theoretical framework emerged that achieved much greater predictive performance: plate tectonics (see Franklin 2012d). The predictive virtues approach delivers the appropriate verdict in this case: Hess's theory should not have been strongly confirmed despite its highly surprising novel predictive success, because its predictive virtues were not very impressive (and were to remain so despite attempts at improvement).

A comparison of these two heterogeneous cases provides some preliminary evidence that the predictive virtues approach has discriminatory capabilities to demarcate between cases where more or less realist commitment is warranted. We can extrapolate a bit further. A pertinent issue we have identified with the anti-realist examples of novelly successful but false theories (e.g. Carman & Díez 2015; Tulodziecki 2017; Rossetter 2018) is that in addition to their novel success, these theories also make multiple false predictions. The fundamental problem with these examples, I suggest, is that they showcase theories whose predictive consequences had only been tested in a rather narrow range of circumstances relative to the domain to which they were applied. Novelty-based realism ends up recommending commitment to these theories, because it elevates novel predictive success as the best criterion for realist commitment in science. However, this is not a consequence of a predictive virtues based defense of scientific realism. For a realist who values predictive virtues instead of novel predictive success, the failures of scientific theories at the frontier of science do not come as a surprise. Predictive flukes can be expected, given the contingent relationship between novelty and truth. False theories may have some true consequences, which sometimes end up as novel successes due to contingent accidents, but these theories also have a number of false consequences, in which their fundamental problems are revealed. A realist who commits to theories based on their predictive virtues will want to wait for more compelling evidence than novel success. Theories that demonstrate veridicality, specificity, scope, and counterfactual depth are more compelling than theories that merely achieve surprising novel success.

Admittedly, our discussion still remains on a vague level. For one, similarly to surprising novel success, predictive virtues also clearly come in degrees. When can we make the determination that the predictive performance of the theory is strong enough to warrant realist commitment (cf. Stanford 2009, p. 384)? Furthermore, are we not still subject to a version of the PMI? Namely, could scientists at previous times have considered the predictive performance of some of their best theories strong enough for realist commitment, and being mistaken, the same fate may yet await us with our current best theories? In what follows, I consider how a predictive virtue based realist could pursue answers to these problems, and perhaps achieve some advances on the realist front.

First, similar to the novelty-based defense of scientific realism, a predictive virtue based defense of scientific realism can be formulated at least in a comparative sense: a predictive virtue based realist holds that theories that demonstrate greater predictive performance than others are more likely

to contain true constituents (cf. Vickers 2013, pp. 196-198). In this, the approach delivers the same result as the competition. However, there is a sense in which it appears to do this better. Whereas we have seen that the surprisingness of the predictions is not associated in a dependable way with the truth of the theory in science, there may be reasons to think that predictive virtues have a better chance. *Prima facie*, it seems that scientific theories tend to become more and more entrenched the stronger their predictive performance becomes. It is not at all clear that there are many examples (if any) of scientific theories that have been able to demonstrate strong predictive performance but have nonetheless turned out predominantly false (see also Fahrbach 2017).[96] A comparative predictive virtue based realism is therefore on the table, and appears prima facie more plausible than the novelty-based approach.

Second, in considering an actual cut-off point for realist commitment, we come back to consider the dynamics between the NMA and the PMI. As far as a precise cut-off point is demanded, there is a sense in which we may ultimately have to come back to a version of the NMA, which others have also acknowledged (see Worrall 2006, p. 51; Fahrbach 2017). If we push far enough the question of why a particular degree of predictive performance is impressive enough from the realist point of view, ultimately what remains is the intuition that certain kind of success simply would not be feasible unless we have a theory that has latched on to something real (e.g. how else could a theory such as general relativity achieve such strong predictive performance if not that it has latched on to reality?).[97] In this, similarly to the novelty-based realist, the predictive virtue based realist may be unable to move the most entrenched anti-realists, who remain unpersuaded by the Miracle intuition or argument (see also Psillos 2011).

There is, however, a new avenue that the predictive virtues approach opens up in a response to the PMI (and in support of the NMA). A new advantage arises from the observation that all scientific theories have a

[96] Arguably, one exception could be Newton's theory, which still appears highly predictive (at least as far as it is not compared to general relativity). I indicated earlier that an advocate of the predictive virtues approach should accept selectivity in confirmation, as it is clear that there can be predictively irrelevant constituents in scientific theories (see section 6.2). Whether or not Newton's theory constitutes a counterexample thus depends on the extent to which the predictions of Newton's theory are based on false rather than true theoretical constituents. If the predictions of Newton's theory are predominantly based on true constituents, it does not constitute a counterexample. Conducting an investigation into the theoretical posits of Newton's system is well beyond the scope of this study; for now, we may simply flag this as an important question in developing the predictive virtue based realist position. (I thank Robert Northcott for raising this issue.)

[97] Here is a way to put the Miracle intuition specifically in terms of predictive virtues. Scientists who have attacked conspiracy theories about the moon landings have sometimes quipped that once you consider all of what it would take to fake the moon landings, including convincingly misleading the competing superpower of the Soviet Union, faking material from the moon, faking videos from the moon, etc., it is easier to just go to the moon. I suggest that something similar is the case with generating a theory that can achieve high predictive performance in a particular domain: in order to produce such a theory, it is easier to just represent what is actually real rather than come up with fabrications that somehow miraculously end up getting the right results.

particular domain to which they are applied (and thus they have certain scope that is required from them). The domain of the theory sets in a rather natural way *limits* to the degree to which theories need to be tested against the evidence: scientific theories can be tested to more or less degrees of comprehensiveness within their domain. With any theory, scientists are always concerned with deriving relevant consequences from the theory, and testing whether observations agree with these consequences. The more such consequences are tested, the less there is left in the world to show that the theory is yet wrong in some respect. At a certain point, a theory may be so thoroughly tested that no relevant tests await that scientists are still looking to perform. I propose that the criterion of *comprehensiveness*, i.e. how thoroughly the theory has been tested in its domain, could be developed as a cut-off point to resist the PMI. It may be that after certain comprehensiveness of testing, no scientific theories have been abandoned in the history of science. Exploring this possibility warrants further research in the future.[98]

Third, unlike in the case of novelty-based realism, the selectionist challenge that we developed in the previous chapter does not apply in the case of predictive virtues. Central to the selectionist response was the idea that *novel* success can be explained by multiple (interdependent) attempts. Isolated surprises can very well arise in the absence of truth or reliability when enough attempts are made over a long period of time. However, if we are concerned specifically with evaluating the predictive virtues of the theory itself, the selectionist challenge does not arise. From the point of view of predictive virtues, what is impressive and in need of explanation is the predictive performance of the theory. Scientists can take all the time in the world to develop a predictive theory about a particular domain, and make as many attempts in the process as they like. If what remains at the end of the process is one theory that is superior to all the others in its predictive virtues, the predictive virtue based realist recommends that the theory be rewarded based on its own merits (cf. Worrall 2011).[99]

Finally, there is one more way in which the predictive virtues approach can sharpen or improve current approaches to the defense of scientific realism. Recently, it has become popular to distinguish between two types of arguments for scientific realism: global 'wholesale' arguments and local 'retail' arguments (see Magnus & Callender 2004). The former type of arguments such as the traditional No Miracles argument provide conclusions about most or all of science, while local arguments are targeted on particular theories, entities, or fields and subfields. The predictive virtues approach applies naturally via what is now the more popular local route: to evaluate

[98] One possibility in defining the appropriate degree of comprehensiveness is that it is settled by the scientists, i.e. it depends on what kind of tests *scientists* themselves want to perform (and are perhaps waiting to perform).

[99] In terms of various underdetermination challenges, the response is similar: in order for the underdetermination challenge to arise, we need an actual challenger to the highly performing predictive theory (and such challenges have not been presented to current best theories).

what kind of predictive performance is impressive in a particular context, we must examine it in the appropriate context of its domain. Despite the local applicability, there is, however, a universal *procedure* to applying the predictive virtue based strategy: in all cases, we are exploring they ways in which scientific theories improve our predictive handle on the world. In this, the universal dimensions of predictivity – veridicality, specificity, scope, and counterfactual depth – also retain a basic unity in the defense of scientific realism. The more we learn about realism in particular contexts, a global argument for scientific realism could also ultimately emerge, where the substantial truth of the best scientific theories is revealed in the overall predictive power they provide over the world.

6.5 PREDICTIVE VERSUS EXPLANATORY VIRTUES

Before we move to consider outstanding issues and potential objections, it is useful to further evaluate the predictive virtues approach in relation to some adjacent debates in the philosophy of science. First, scientific explanation. Many of the predictive virtues that have been introduced here could also be invoked in categorizing *explanatory virtues* (see, for example, Ylikoski & Kuorikoski [2010], who invoke 'precision' and 'factual accuracy' among five dimensions of explanatory power, embedded in a contrastive-counterfactual approach to explanation). What is the relationship between predictive virtues and explanatory virtues, and why is a separate notion of predictive virtues needed in addition to explanatory virtues?

First, it should be pointed out that the fact that there is overlap between explanatory and predictive virtues is unsurprising: prediction and explanation have for long been recognized as closely related concepts in the philosophy of science (see Douglas 2009). In a recent paper, aptly titled "Explanation, Prediction, and Causality: Three Sides of the Same Coin?", Watts et al. (2018) remind us that whenever we are talking about causal explanation in science, we are *always* making predictions, whether they are explicitly stated or not. Any explanation that cites X as the cause of Y is in effect making predictions about how changes in X would impact Y. A certain kind of predictivity is thus constitutive of all causal explanations, and the fact that these concepts are closely related is not unexpected.

However, despite the close relationship, there are also important differences, which indicate the need for a separate notion of predictive virtues. First, unlike prediction and predictive virtues, explanation also includes a dimension of cognitive salience or comprehensibility, where proper explanations should cite cognitively accessible causes or mechanisms for the explanandum (see Kuorikoski & Ylikoski 2010). Cognitive salience is undoubtedly a desirable demand as such, but the problem is that it also leaves important parts of modern science out. Scientists are increasingly exploring the use of new methods such as machine learning algorithms with the purpose

of achieving more predictive science (e.g. Yarkoni & Westfall 2017). These methods often produce models that are not cognitively comprehensible, but they can nonetheless improve predictive success. It is useful to have a distinct concept of predictive virtues that captures the various ways in which such models can be successful for scientific purposes without providing explanations.

Another reason for introducing a separate notion of predictive virtues is that predictivity and explanatoriness, as distinct sets of properties, diverge in both their epistemic and pragmatic implications. A plausible case can be made that all of the predictive virtues we have introduced in this study are important from the confirmatory point of view (see section 6.2). However, explanatoriness includes dimensions such as cognitive salience, which may be less relevant to theory confirmation (we can ask, does cognitive salience make a hypothesis more likely to be true about the mind-independent world?). Some, in fact, have argued that explanatoriness is altogether irrelevant from the confirmatory point of view (see Roche & Sober 2013; for further discussion, see Roche & Sober 2014, 2017; McCain & Poston 2014; Climenhaga 2017). There is also a sense in which predictivity appears to be more fundamental than explanatoriness to confirmation. Northcott (2017) argued recently that there is no explanation without prediction in science: after all, whenever we attempt to provide causal explanations, we are also committing ourselves to predictions, and only the success of the predictions can provide epistemic support for the explanations (cf. Watts et al. 2018).[100] Further studies are needed to explore and compare the epistemic and pragmatic implications of both predictivity and explanatoriness. For example, a particularly interesting question would be to explore when (and why), if ever, does explanatoriness contribute to theory confirmation over and above predictivity?

6.6 PREDICTIVE VIRTUES AND SCIENTIFIC PROGRESS

Another adjacent topic in the philosophy of science concerns the idea of scientific progress, i.e. whether there are improvements or advancements in scientific representations of the world. How do we best conceptualize and measure scientific progress (assuming there is such a thing)?

Multiple theories have been advanced in the philosophy of science (see Niiniluoto 2019 for a review). These include interpreting progress in terms of truth, truthlikeness (e.g. Niiniluoto 2017), knowledge (e.g. Bird 2007), and understanding (e.g. Dellsén 2016). The details of this debate are beyond the scope of the current study, but it is useful to briefly observe that viewing scientific progress in terms of the predictive virtues approach offers a rather natural and attractive alternative. It is uncontroversial that science has

[100] Northcott (2017) uses the term 'prediction' roughly in the sense of what we have called logical prediction, i.e. it covers both forecasting (predicting something in the future) and retrodiction (after-the-fact derivations from theory).

achieved improvements with all of the predictive virtues that have been introduced in this study: current scientific theories have far outperformed previous theories in their predictive performance (see Fahrbach 2017). As science has advanced over decades and centuries, we have gradually been presented with theories that are more accurate, more precise, more widely applicable, and enable more counterfactual inferences about the world than their predecessors. The predictive virtues approach provides a straightforward way to interpret scientific progress in terms of these advancements that are readily visible in scientific practice.

More specifically, there are at least three potential advantages that the new approach brings to the discussion. First, the predictive virtues approach enables comparisons between competing theories and theoretical frameworks so that we only have to consider the theory and the evidence as such (along with relevant background knowledge). To evaluate whether there has been progress moving from one theory or theoretical framework to another, we can compare the predictive virtues of the two theories. If one theory displays greater performance in terms of multiple predictive virtues (without falling behind in others), we can say that it has achieved progress over the other theory in scientific representation of the world. This is not to say that such comparisons must always be easy, or that progress itself is always easy to determine. Rather, the advantage of the predictive virtues approach is that it offers a relatively straightforward procedure for how to evaluate scientific progress, including determining whether there is any clear sense in which it has occurred. Another benefit of the approach is that it brings the potential of measuring scientific progress in a more nuanced way. Other theorists have incorporated the ability to derive correct predictions in their account of scientific progress (e.g. Dellsén 2016). However, further distinguishing between different types of predictive successes or virtues enables capturing this type of progress in a more nuanced and comprehensive way. In this, the predictive virtues approach could at least complement other, extant approaches to scientific progress.

A final advantage of the predictive virtues approach is that it achieves a certain kind of unity in representing various aspects of scientific progress. Scientific theories and models often contain features that may be difficult to account for from the point of view of other notions such as truth or knowledge. One example are idealizations (see Weisberg 2007). Scientific theories and models often contain idealizations that are strictly false and hence do not count as such for either truth or knowledge. However, the idealizations may nonetheless be useful for some scientific purpose; typically, prediction. The predictive virtues approach incorporates progress that is achieved based on idealizations: what counts is whether or not predictive performance increases in consequence of introducing the idealizations. Another example is prediction without explanation (and understanding). As the amount of data has increased in the world, scientists have increasingly begun to develop more complex models that do not seek to provide explanations of their target system

(e.g. Yarkoni & Westfall 2017). The explicit purpose of these models is prediction. The predictive virtues approach introduces a natural way to classify and evaluate the different ways in which scientific advancements are made with complex predictive models. Finally, as discussed in the previous section, there is also a sense in which in so far as science does achieve explanations (and truth and knowledge), this already entails that we also have predictive performance. If this is correct, prediction is needed even for these kind of scientific advancements. In this, the predictive virtues approach potentially unifies multiple different ways in which scientific progress is achieved, constituting perhaps a more central criterion according to which such progress should be evaluated. A more thorough exploration and development of this idea must also wait for another occasion.

6.7 QUESTIONS, CLARIFICATIONS, AND OBJECTIONS

As we are approaching the end of our investigation, we should return now to the various findings, issues, and concerns that we have uncovered in the course of this study. How does the predictive virtues approach incorporate these findings, and handle outstanding issues? What kind of objections could be raised against the new approach?

1) We began our study of the prediction versus accommodation problem by citing three desiderata for an adequate predictivist account: i) the account should be objective rather than subjective or psychologistic, ii) it should agree in important cases with the confirmatory evaluations of scientists, and iii) it should have a clear epistemological rationale (see Mayo 2014, p. 80). Does the predictive virtues approach meet these standards?

I suggest that the answer is 'yes' on all counts. First, the new approach is not subjective or psychologistic: it explicitly rejects that contingent factors related to the theorist's use of evidence should have a central role in confirmation theory. Second, the new approach implies confirmatory verdicts that agree with scientific judgments in multiple cases, cases that have otherwise been difficult to incorporate in previous predictivist approaches. For example, the new approach implies that Hess's theory of continental drift should not have been strongly confirmed despite its surprising novel predictive success, but it also says that general relativity should have due to its veridicality, specificity, scope, and counterfactual depth. Finally, the new approach has a clear epistemological rationale. Central to scientific practice is the idea that science is in the business of looking for patterns, regularities, and mechanisms in the world, and seeking to provide accurate and comprehensive theories about those patterns, regularities, and mechanisms. Each of the predictive virtues introduced here measures in distinct ways the success of scientific theories in achieving this goal.

2) Another central problem we have sought to resolve is the "paradox of predictivism" (see Barnes 2008). How is it that it simultaneously

seems that novel predictions are the best evidence for a scientific theory, but novelty depends on contingent accidents which should have no place in confirmation theory?

I suggest that we can resolve this paradox by recognizing that prediction matters for confirmation in the *logical* sense. What we are after in science are theories that are able to achieve strong predictive performance in relationship to the world. The contingent dimension of novelty itself is not important, and it is properly seen as an accident. It is true that in some cases, a predictively powerful theory happens to *also* generate an impressive novel prediction. This happened, for example, with general relativity and the bending of light. In these cases, it may have been attractive to think that the fact that the phenomenon is new is what is important from the epistemic point of view. (A number of people have commented on the *psychological* value of novel predictions: new discoveries are exiting, and may raise a lot of interest as such [e.g. Ne'eman & Kirsch 1986, p. 202; Brush 1989; Scerri & Worrall 2001, p. 426].) However, I submit that what actually counts for theory confirmation is the theory's powerful demonstration of predictive virtues. The bending of light prediction, as a logical prediction of general relativity, confirms this theory because this prediction demonstrates powerful specificity and counterfactual depth in that theory. General relativity would have been none the worse from the confirmatory point of view if it had been developed based on knowledge about the bending of light. This claim is supported by the finding that in other cases in science impressive novel successes have come *without* compelling predictive virtues, and in these cases they appear to count for much less in the scientific community (e.g. Hess's mantle-based theory of drift). This also dovetails the results of Brush (1994, 1995), who found that scientific theories are sometimes rejected despite their novel success, and accepted without novel successes. Moreover, Brush (1989) found that general relativity in specific was not accepted by physicists based on its novel success; what counted was its empirical performance relative to alternative theories.[101]

3) In chapter 4, it was found that there are multiple methodological issues that potentially affect the degree to which we should have confidence in a scientific hypothesis. These include p-hacking, overfitting, small sample sizes, hypothesis hunting, fraud, etc. Overall, it was argued that the methodological issues do not show that novel prediction should be preferred to accommodation in scientific practice. However, these issues do clearly have potential epistemic implications in themselves. What is the relationship of the predictive virtues approach to the various methodological issues in scientific practice?

The new approach identifies a goal of scientific practice: generate a predictively virtuous theory. Methodological rules and appropriate methodological practices, I suggest, matter because of the results they

[101] Brush (1989, p. 1126) also reveals that Sir Joseph Thomson, with whom we began this study, found general relativity at the time "too abstruse to be acceptable" (despite the fact that he highlighted its novel success).

produce: bad practices lead to theories that fail to achieve adequate predictive performance (for example, these practices produce theories that fail under the first attempts at replication). The role of the predictive virtues approach is to set and clarify standards for the results that scientific methodologies should produce: theories that are predictively virtuous and can so withstand the test of reality.

The predictive virtues approach does also have implications that concern methodological practices and issues individually. For example, the predictive virtues approach does not prohibit HARKing, or hold that there is a general confirmatory penalty associated with HARKing. In other words, according to this approach, there is nothing in principle wrong with constructing hypotheses based on the results of statistical experiments. However, the predictive virtues approach, by virtue of demanding certain kind of predictive performance from the theory, does set constraints on this practice. Constructing hypotheses after-the-fact is acceptable to the extent that we have evidence for the predictivity of the new hypothesis. Using a small sample to construct a hypothesis that appears spurious in the light of background knowledge does not warrant much (if any) confirmation. If the sample is large and/or there are further evidence or reasons that support the hypothesis, the degree of confirmation is higher. Similarly, the predictive virtues approach sets limits on p-hacking. The predictive virtues approach demands that scientific theories make accurate rather than inaccurate predictions, so p-hacking is an issue to the extent it hides actual predictive inaccuracies in the tested hypothesis.

Finally, rules about statistical model selection methods are also naturally backed up by the predictive virtues approach. The purpose of model selection methods such as the AIC and cross-validation is to ensure that we pick the *predictively* most powerful model out of multiple possibilities. For example, the novelty-rule that is utilized in cross-validation is applied specifically for the purpose of selecting the model that is likely to have the best long run predictive accuracy. In other words, it is used to select the model whose overall predictivity is maximal, rather than, say, the model that represents the data best in one particular dataset.

4) We found that novel predictions can sometimes become relevant to confirmation in certain limited circumstances. This happens when we do not have access to evaluate the theory and the evidence as such (e.g. in cases of epistemic opacity and perhaps certain cases of overfitting). How does the predictive virtues approach incorporate this finding?

In this respect, the predictive virtues approach remains similar to Worrall's (2014) original logical predictivist theory, in that the confirmation provided by novel prediction in these situations follows as a special case from the overall theory. If we cannot evaluate the predictive virtues of the theory directly, novel predictive success can provide some evidence about underlying logical predictivity. Any temporally novel or use-novel prediction that is derived from a theory is also a logical prediction of that theory, so by observing

such a prediction, we can obtain some evidence about potential logical predictivity even if we cannot evaluate the theory directly.[102] However, crucially, the predictive virtues approach does not place any special emphasis on these novel predictions. They count only to the extent that they reveal us something about underlying logical predictivity. So, for example, in cases such as epistemic opacity, the confirmation provided by novel predictive success remains extremely low, because it does not enable us to evaluate the predictive virtues of the model in any reliable way. The predictive virtues approach delivers the appropriate verdict that scientific confirmation ultimately concerns the theory and its relationship with the evidence, and this is best evaluated if we have the theory and the evidence explicitly on the table (cf. Worrall 2014, p. 58).

5) One more outstanding issue concerns the layman perspective to theory evaluation. We have thoroughly discussed the prediction versus accommodation issue from the scientific point of view, but what do our results show about the layman perspective?

All of our results in Chapters 4, 5, and 6 concern the layman evaluator as well as scientific evaluators: novel predictive success is no more reliable epistemic indicator in theory evaluation for the layman that it is for the scientist. The overall epistemic recommendation for the layman is the same as for scientific evaluators: we as laymen should be concerned with the predictivity of the scientific theory rather than surprising novel success. As far as we cannot evaluate predictivity ourselves, the better epistemic strategy seem to be to go by the judgment of the people who have the appropriate access to evaluate it (i.e. scientists) rather than put our faith in unreliable indicators such as novel predictive success (see also Brush 1994, 1995; Harker 2008).

Let us next consider some possible objections to the new approach:

6) Some might question whether the new approach is needed in the context of logical theories of confirmation. After all, we already have a prominent logical theory of confirmation: Bayesianism. Why do we need the predictive virtues approach in addition to standard Bayesianism?

The new approach brings certain advantages that can be useful in further developing logical confirmation theories. First, the Bayesian calculus is notoriously silent when it comes to questions such as how do we demarcate between scientific and unscientific hypotheses, where does the prior degree of confidence in the hypothesis come from, what makes particular type of evidence more confirming, etc. The predictive virtues approach can supplement the more austere logical theories of confirmation in this regard. It

[102] It should be observed that not all temporally novel or use-novel predictions are necessarily always logical predictions that are derived from theory. It is, after all, possible to simply state a guess about the future. In cases where we do not know the contents of the theory, a temporally novel or use-novel prediction may provide some evidence about potential logical predictivity of an underlying theory, but this evidence is appropriately seen as weak, at least as far as there are no further reasons to have confidence in the underlying theory (e.g. there are further demonstrations of actual predictive virtues such as accurate, specific, counterfactually deep predictions).

provides criteria for distinguishing good theories and good evidence from bad ones, and thus enables a more precise theory of confirmation that may more closely capture the logic of scientific confirmation. For example, referring to predictive virtues, we could perhaps set priors in a more objective way based on the predictive virtues of the theory. Similarly, the degree of confirmation provided by any individual prediction could be measured more precisely based on its specificity and/or counterfactual depth (using Bayesian conditionalization). In this, despite bringing something new to the table, the predictive virtues approach does not take away the old. It could act simply as a supplement to Bayesianism, addressing some of its shortcomings. Second, the predictive virtues approach also provides solutions to multiple puzzles about the confirmatory role of predictions in science. In other words, it has explanatory power over epistemic practices in science, over and above strict Bayesianism.[103]

7) The new approach makes logical prediction or predictivity into a central concept for scientific confirmation. Is everything there is to theory confirmation about prediction and predictivity?

Scientific theories are ultimately evaluated based on their contact with empirical reality (see also Northcott 2017). The new approach attempts to capture the various ways in which theories are empirically successful in the appropriate sense. In this regard, I would endorse the claim that a considerable part of scientific confirmation does come down to logical prediction and predictivity. However, this is not to deny that there are no other epistemic factors that are also relevant to theory confirmation. Background knowledge still counts as per usual; for example, theories can gain further support if they fit other, already confirmed scientific theories. We should also investigate other potential sources of epistemic support that have been debated in the philosophy of science (e.g. robustness, explanatory unification), and evaluate their relationship to predictive virtues.

From the converse point of view, it should also be emphasized that the new approach does not hold that all theory-evidence relationships are automatically 'predictive' in the appropriate sense. On the contrary, many types of relationships between the theory and empirical results are ruled as non-predictive in the new approach. For example, there are various pseudo-explanations that seem intuitively unsupported by the evidence. Schurz (2014, p. 88) mentions "a flying spaghetti monster has brought it about that grass is green" as a pseudo-explanation for the fact that grass is green. This hypothesis fails to demonstrate just about any predictive virtues in relation to the evidence (apart from the veridicality of the trivial logical consequence that grass is green, already cited in the hypothesis itself), so the hypothesis is ruled out as non-predictive in our sense and thus not confirmed by this evidence.[104]

[103] In this, the predictive virtues approach also has underlying predictive power: it provides rather precise predictions about what kind of theories should and should not enjoy support among scientists.

[104] Pseudo-explanations provide one way to test philosophical theories of scientific confirmation. The predictive virtues we have set, particularly combined with the selectivist criterion of confirmation,

Pure logical accommodations (i.e. free parameters that can be adjusted to any evidence from reality) are similarly ruled out as irrelevant to theory confirmation due to the fact that they fail to meet the criterion of predictive specificity.

8) The predictive virtues approach could end up increasing the complexity of confirmation theory. Is it perhaps already *too* complex?

The predictive virtues approach recommends that we evaluate scientific predictions from a more holistic perspective, where both individual predictions and the predictivity of the theory from which they are derived count. This adds some degree of complexity. However, at the same time, by denying that contingent factors are important to confirmation, the new approach also reduces complexity in a meaningful way. In the end, the predictive virtues approach could be simpler than novelty-based approaches to prediction and confirmation. With predictive virtues, confirmatory evaluations involve only the theory and the evidence as such (along with relevant background knowledge). To evaluate the predictivity of a scientific theory, we start by identifying the domain or target to which the theory is applied. Next, we study to what extent the theory makes predictions within that domain, and whether or not these predictions have turned out accurate. To further evaluate the degree of confirmation provided by individual predictions, we examine their specificity and counterfactual depth. These evaluations can be made in a relatively straightforward manner, as we have seen with different scientific examples in the course of this chapter. Further theoretical development on quantitative measures of the degree of confirmation provided by particular predictions (e.g. a measure of specificity) could also ultimately make these evaluations rather precise.

6.8 SUMMARY

In this chapter, a new logical approach to predictivism and scientific confirmation was developed. It was argued that previous predictivist approaches have focused on the "wrong" kind of predictions in science, as well as understood scientific prediction altogether too narrowly. In scientific practice, the fundamental object or unit that makes predictions is the scientific theory or model, and the evaluation of predictions and their epistemic implications must focus on the performance of the theory or the model itself. Four predictive virtues that measure the predictivity of the theory were introduced: veridicality, specificity, scope, and counterfactual depth. It was

can provide robust limits to rule out such pseudo-explanations. Pseudo-explanations such as Schurz's spaghetti monster example usually connect to trivially unspecific empirical facts, they introduce inflated theoretical constructs that are not needed for the empirical predictions of the theory, and they typically have no counterfactual depth at all (or, if they do, their predictions are wildly inaccurate; e.g. such as predictions of astrology). Further investigations of pseudo-explanations, and the ability of the predictive virtues approach to deal with them, are called for in the future to test and further develop the approach.

argued that what counts for scientific confirmation is the predictive performance of the theory itself, as revealed in its predictive virtues.

The new predictive virtues approach was evaluated against multiple puzzles and problems that challenge predictivist theories of confirmation. It was argued that it can handle the puzzles better than the competition. First, the new approach appropriately distinguishes between cases where evidence used in the construction of the theory either counts or does not count for confirmation. What counts is that the theory demonstrates predictive virtues in relationship to the evidence, not which evidence was used or needed to be used at which point in the construction of the theory. Second, the new approach can distinguish between cases where more or less realist commitment is warranted: what matters are the predictive virtues of the theory, not the surprisingness of its predictions. Finally, the new approach fits naturally into adjacent debates in the philosophy of science, carving out a unique perspective to topics such as explanation and scientific progress. In this, it may constitute a fruitful new opening to thinking about the various benefits, epistemic and otherwise, of prediction in science.

7 CONCLUSION

This study investigated the epistemic role and value of prediction in science. *Predictivists* have typically defended the view that *novel* predictions, i.e. empirical consequences of the theory that were not used in its construction, have special epistemic value. Predictivists maintain that novel predictions confirm the theory more strongly than accommodations and they provide the grounds for realist commitment to scientific theories. In the defense of these claims, two strategies have been employed. One is negative: it argues that novel prediction should be preferred because accommodation is associated with negative epistemic consequences. The other is positive: it argues that novel prediction is better because only novel success calls for a realist explanation. Both strategies were evaluated and rejected in this study. When other epistemic factors and the problems of novel prediction are taken into account, there is no general or important advantage to novel prediction due to the (potential) problems of accommodation. When the competitive context of scientific practice where novel predictions are pursued is taken into account, it also turns out that there is no need for a realist explanation in cases of novel success. In accordance, novelty-based predictivism was rejected as a viable approach to the epistemic role and value of prediction in science in this study.

In contrast to the novelty-based approach, another approach to the epistemic role of prediction in science considers predictions from a logical point of view. The logical approach to predictivism argues that scientific predictions should be understood from the point of view of the theory, and in terms of how it stands in relation to empirical results in the world. This view has been defended prominently by John Worrall (2006, 2009, 2014), who argues that scientific theories are confirmed in cases where they achieve logical predictive success, i.e. when they generate independently testable and natural consequences. Worrall's confirmation theory was evaluated and ultimately rejected, because it inherits many of the problems of contingent novel prediction. However, it was argued that the logical approach itself is promising and worthy of further development.

To this effect, a new *predictive virtues approach* on the epistemic role and value of prediction in science was introduced. The new approach subscribes to the logical approach in that it takes the logical predictions or predictivity of the theory as what is important from the epistemic point of view. However, it was argued that the predictivity of the theory is a more nuanced and multifaceted property than has been previously recognized in the philosophical literature. There are at least four distinct ways in which scientific theories can be more or less predictive with regard to the world: in terms of their veridicality, specificity, scope, and counterfactual depth. These virtues or dimensions form the basis of the new predictive virtues approach, which argues that scientific theories should be evaluated based on their predictive

performance. It was argued that the predictive virtues approach can account for multiple anomalies that plague previous predictivist approaches, and in this it makes advances in the discussion.

The study has multiple consequences in the philosophy of science, and it raises a host of further questions and topics of study. First, the predictive virtues themselves require further investigation. More precise quantitative measures ought to be developed that measure the confirmatory impact of single (specific or counterfactually deep) predictions, and the predictivity of scientific theories needs to be studied more closely in the domains in which they are applied. The relationships between the virtues also require further study; for example, some of them may be in some sense hierarchical (see fn. 90). Furthermore, the question of the epistemic and pragmatic value of distinct predictive, as well as explanatory, virtues should be further investigated. An interesting line of research would be to explore when, if ever, does explanatoriness contribute to confirmation over and above predictivity.

Second, the results of the study imply that both the discussion on prediction vs. accommodation and scientific realism vs. anti-realism need some adjustment. With regard to the novel prediction versus accommodation distinction, the results of the study are largely negative: this distinction does not appear to be a particularly important one to confirmation in scientific practice. The distinction may count in some circumstances (e.g. epistemic opaqueness), but in terms of importance, it appears to be more peripheral rather than central when it comes to what matters epistemically in science. It is recommended that in place of debating the issue of novel prediction vs. accommodation, it may be better to start considering the various ways in which scientific theories can achieve predictive successes and advances. There may be *a* distinction between novelty and accommodation, i.e. between scientists using and not using evidence, but more fundamental epistemic patterns related to scientific prediction and confirmation await the attention of philosophers of science.

In terms of scientific realism, the results of the study imply that the novelty-based defense of scientific realism fails. Novel predictive successes do not constitute a dependable basis for realist commitment to the working constituents of novelly successful theories. With regard to this specific argument, which has been called "the Ultimate Argument" for scientific realism (see Psillos 2006), the anti-realist thus has the upper hand. However, it was suggested that the predictive virtues approach might offer a new way forward for the scientific realist. Perhaps, a better way to allocate realist commitment is to go by the predictive performance of the theory itself rather than contingent accidents related to who happened to use some (surprising) evidence or not. To this effect, more detailed case studies that investigate the various predictive virtues and their potential realist implications are called for. Studies of exemplars of highly performing predictive theories could also be conducted to further clarify and develop the predictive virtues approach (see Saatsi 2017).

The study also has some consequences when it comes to scientific practice, at least in terms of a few (modest) epistemic recommendations. First, the results of the study speak particularly to the value of replication studies in scientific practice (cf. Bakker et al. 2012), as well as to the use of more data (cf. Yarkoni & Westfall 2017). Science at the frontier is in many ways uncertain, whether we are talking about novel predictions or accommodations. There are multiple issues that concern new and isolated studies that imply that the actual predictive performance of the theory at hand may not be what it initially seems. Confidence in the theory increases the more it is tested against evidence from the world, and the clearer it becomes to what degree the theory is predictive in various ways. Second, the results of the study also speak to the value of transparency in scientific research (cf. Aguinis, Ramani & Alabduljader 2018). The more transparently the contents of the theory, the evidence, and methodological choices are reported, the more dependable epistemic evaluations can be made, even if we are yet considering the results of a single study.

Finally, one further consequence of the study concerns more broadly the nature of the scientific enterprise and how it is understood from the philosophical point of view. Ever since Popper, a prevalent view of science has been that science progresses through 'bold' conjectures (see Popper 1963). Good scientific practice is to come up with bold new hypotheses, and then derive novel consequences from those hypotheses to test them against the world. The hypothetico-deductive model of scientific research has been widely influential in scientific practice, guiding the methodology of several generations of scientists (see, for example, Kerr 1998; Leung 2011; Nosek et al. 2018). Opposed to this view, there is another, less widely appreciated idea according to which there is another proper way to conduct science: collect evidence, and use it to construct theories and hypotheses (see, for example, Brush 2015; Scerri 2016). From this point of view, it is also possible to engage in good scientific practice by collecting more and more evidence and seeking to come up with good models and theories about that evidence. This view is becoming more and more relevant in the contemporary push to make science more predictive through the use of more data and better methods in learning from the patterns in the data (e.g. Yarkoni & Westfall 2017). The alternative, data-driven view turns the conjecture-based view on its head: instead of starting with theory, we can also start with the evidence, and see where it leads us. The results of this study imply that neither the data-driven nor the conjecture-based view provides the sole, most accurate account of good scientific practice: both constitute potential and significant ways of advancing scientific research. In this, a "new," more inclusive picture of science is promoted, where building theories based on the evidence can be just as good, and sometimes even better, as proceeding through bold conjectures.

REFERENCES

Aguinis, H., Ramani, R.S. & Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, 12(1), 83–110. https://doi.org/10.5465/annals.2016.0011.

Akaike, H. (1973). Information Theory as an Extension of the Maximum Likelihood Principle. In B.N. Petrov & F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267–281.

Akeroyd, F.M. (2003). Prediction and the Periodic Table: a response to Scerri and Worrall. *Journal for General Philosophy of Science*, 34(2), 337–355. http://www.jstor.org/stable/25171262.

Alai, M. (2014). Novel Predictions and the No Miracle Argument. *Erkenntnis*, 79, 297–326. https://doi.org/10.1007/s10670-013-9495-7.

Alai, M. (2016). The No Miracle Argument and Strong Predictivism Versus Barnes. In L. Magnani, & C. Casadio (eds.), *Model-Based Reasoning in Science and Technology. Logical, Epistemological, and Cognitive Issues. Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer International Publishing, 541–556. https://doi.org/10.1007/978-3-319-38983-7_30.

Alai, M. (2017). Resisting the historical objections to realism: Is Doppelt's a viable solution? *Synthese*, 194(9), 3267–3290. http://dx.doi.org/10.1007/s11229-016-1087-z.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *Leadership Quarterly*, 28, 5–21. https://doi.org/10.1016/j.leaqua.2017.01.006.

Arlot, S. & Celisse, A. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, 4, 40–79. https://doi.org/10.1214/09-SS054.

Asay, J. (2019). Going local: a defense of methodological localism about scientific realism. *Synthese*, 196, 587–609. https://doi.org/10.1007/s11229-016-1072-6.

Bakker, M., van Dijk, A. & Wicherts, J.M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, 7(6), 543–554. https://doi.org/10.1177/1745691612459060.

Barnes, E.C. (1996). Social Predictivism. *Erkenntnis*, 45, 69–89. https://doi.org/10.1007/BF00226371.

Barnes, E.C. (2002a). Neither Truth Nor Empirical Adequacy Explain Novel Success. *Australasian Journal of Philosophy*, 80(4), 418–431. https://doi.org/10.1080/713659528.

Barnes, E.C. (2002b). The Miraculous Choice Argument for Realism. *Philosophical Studies*, 111, 97–120. https://doi.org/10.1023/A:1021204812809.

Barnes, E.C. (2008). *The Paradox of Predictivism*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511487330.

Barnes, E.C. (2014). The roots of predictivism. *Studies in History and Philosophy of Science Part A*, 45, 46–53. https://doi.org/10.1016/j.shpsa.2013.10.002.

Barnes, E.C. (2018). Prediction versus Accommodation. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Fall 2018 Edition)*.

URL = <https://plato.stanford.edu/archives/fall2018/entries/prediction-accommodation/>.

Baumberger, C., Knutti, R. & Hadorn, G.H. (2017). Building confidence in climate model projections: an analysis of inferences from fit. *Wiley Interdisciplinary Reviews: Climate Change*, 8(3), e454. https://doi.org/10.1002/wcc.454.

Bird, A. (2007). What is scientific progress? *Noûs*, 41(1), 64–89. https://doi.org/10.1111/j.1468-0068.2007.00638.x.

Boge, F.J. (2020). An argument against global no miracles arguments. *Synthese*, 197, 4341–4363. https://doi.org/10.1007/s11229-018-01925-9.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophy of Science*, 97(3), 303–352. https://doi.org/10.2307/2185445.

Bosco, F.A., Aguinis, H., Field, J.G., Pierce, C.A. & Dalton, D.R. (2016). HARKing's Threat to Organizational Research: Evidence From Primary and Meta-Analytic Sources. *Personnel Psychology*, 69(3), 709–750. https://doi.org/10.1111/peps.12111.

Boyce, K. (2018). The Coincidentalist Reply to the No-Miracles Argument. *Erkenntnis*, 83, 929–946. https://doi.org/10.1007/s10670-017-9920-4.

Brush, S.G. (1989). Prediction and Theory Evaluation: The Case of Light Bending. *Science*, 246(4934), 1124–1129. https://doi.org/10.1126/science.246.4934.1124.

Brush, S.G. (1990). Prediction and Theory Evaluation: Alfvén on Space Plasma Phenomena. *Eos*, 71(2), 19–33. https://doi.org/10.1029/EO071i002p00019.

Brush, S.G. (1993). Prediction and Theory Evaluation: Cosmic Microwaves and the Revival of the Big Bang. *Perspectives on Science*, 1(4), 565–601.

Brush, S.G. (1994). Dynamics of theory change: The role of predictions. *PSA*, 2, 133–145. https://doi.org/10.1086/psaprocbienmeetp.1994.2.192924.

Brush, S.G. (1995). Prediction and Theory Evaluation in Physics and Astronomy. In A.J. Kox & D.M. Siegel (eds.), *No Truth Except in the Details. Essays in Honor of Martin J. Klein*. Dordrecht: Kluwer, 299–318.

Brush, S.G. (1996). The Reception of Mendeleev's Periodic Law in America and Britain. *Isis*, 87(4), 595–628. https://doi.org/10.1086/357649.

Brush, S.G. (2007). Predictivism and the Periodic Table. *Studies in the History and Philosophy of Science Part A*, 38(1), 256–259. https://doi.org/10.1016/j.shpsa.2006.12.007.

Brush, S.G. (2015). *Making 20th Century Science: How Theories Became Knowledge*. Oxford: Oxford University Press.

Carman, C. & Díez, J. (2015). Did Ptolemy make novel predictions? Launching Ptolemaic astronomy into the scientific realism debate. *Studies in History and Philosophy of Science Part A*, 52, 20–34. https://doi.org/10.1016/j.shpsa.2015.04.002.

Chakravartty, A. (2017a). Scientific Realism. In E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Summer 2017 Edition)*.

URL = <https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/>.

Chakravartty, A. (2017b). Reflections on new thinking about scientific realism. *Synthese*, 194, 3379–3392. https://doi.org/10.1007/s11229-017-1514-9.

Climenhaga, N. (2017). How Explanation Guides Confirmation. *Philosophy of Science*, 84(2), 359–368. https://doi.org/10.1086/690723.

Dawid, R. & Hartmann, S. (2018). The no miracles argument without the base rate fallacy. *Synthese*, 195, 4063–4079. https://doi.org/10.1007/s11229-017-1408-x.

Dellsén, F. (2016). Scientific Progress: Knowledge versus Understanding. *Studies in History and Philosophy of Science*, 56, 72–83. https://doi.org/10.1016/j.shpsa.2016.01.003.

Dellsén, F. (forthcoming). An Epistemic Advantage of Accommodation over Prediction. *Philosophers' Imprint*.

Díez, J. (2018). A (Fatal) Trilemma for best theory realism. *European Journal for Philosophy of Science*, 8, 271–291. https://doi.org/10.1007/s13194-017-0185-1.

Doppelt, G. (2005). Empirical Success or Explanatory Success: What Does Current Scientific Realism Need to Explain? *Philosophy of Science*, 72(5), 1076–1087. https://doi.org/10.1086/508958.

Doppelt, G. (2014). Best Theory Scientific Realism. *European Journal for Philosophy of Science*, 4, 271–291. https://doi.org/10.1007/s13194-014-0090-9.

Douglas, H. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, 76(4), 444–463. https://doi.org/10.1086/648111.

Douglas, H. & Magnus, P.D. (2013). State of the Field: Why novel prediction matters. *Studies in History and Philosophy of Science Part A*, 44(4), 580–589. https://doi.org/10.1016/j.shpsa.2013.04.001.

Duhem, P. (1914/1954). *The Aim and Structure of Physical Theory*, trans. from 2nd ed. by P.W. Wiener; originally published as *La Théorie Physique: Son Objet et sa Structure* (Paris: Marcel Riviera & Cie.). Princeton: Princeton University Press.

Gardner, M.R. (1982). Predicting Novel Facts. *The British Journal for the Philosophy of Science*, 33(1), 1–15. https://www.jstor.org/stable/687237.

Giere, R.N. (1984). *Understanding Scientific Reasoning, 2nd edition*. New York: Holt, Rinehart, and Winston.

Fahrbach, L. (2009). The pessimistic meta-induction and the exponential growth of science. In A. Hieke & H. Leitgeb (eds.), *Reduction–Abstraction–Analysis*. Lancaster: Ontos Verlag, 95–112. https://doi.org/10.1515/9783110328875.

Fahrbach, L. (2011a). Theory Change and Degrees of Success. *Philosophy of Science*, 78(5), 1283–1292. https://doi.org/10.1086/662280.

Fahrbach, L. (2011b). How the growth of science ends theory change. *Synthese*, 180, 139–155. https://doi.org/10.1007/s11229-009-9602-0.

Fahrbach, L. (2017). Scientific revolutions and the explosion of scientific evidence. *Synthese*, 194, 5039–5072. https://doi.org/10.1007/s11229-016-1193-y.

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE*, 4(5), e5738. https://doi.org/10.1371/journal.pone.0005738.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. & Rummukainen, M. (2013). Evaluation of climate models. In T.F. Stocker, D. Qin, G.K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P. Midgley (eds.), *Climate Change 2013: The physical science basis. Contribution of*

Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge/New York: Cambridge University Press.

Forbes, C. (2017). A pragmatic, existentialist approach to the scientific realism debate. *Synthese*, 194, 3327–3346. https://doi.org/10.1007/s11229-016-1015-2.

Forsyth, D. & Uyeda, S. (1975). On the Relative Importance of the Driving Forces of Plate Motion. *Geophysical Journal International*, 43(1), 163–200. https://doi.org/10.1111/j.1365-246X.1975.tb00631.x.

Forte, A.M. (2011). Plate Driving Forces. In H.K. Gupta (ed.), *Encyclopedia of Solid Earth Geophysics*. Dordrecht: Springer, 977–983. https://doi.org/10.1007/978-90-481-8702-7.

Frankel, H.R. (2012a). *The Continental Drift Controversy. Volume 1: Wegener and the Early Debate*. New York. Cambridge University Press.

Frankel, H.R. (2012b). *The Continental Drift Controversy. Volume 2: Paleomagnetism and Confirmation of Drift*. New York. Cambridge University Press. https://doi.org/10.1017/CBO9780511843167.

Frankel, H.R. (2012c). *The Continental Drift Controversy. Volume 3: Introduction of Seafloor Spreading*. New York. Cambridge University Press. https://doi.org/10.1017/CBO9781139025416.

Frankel, H.R. (2012d). *The Continental Drift Controversy. Volume 4: Evolution into Plate Tectonics*. New York. Cambridge University Press. https://doi.org/10.1017/CBO9781139095938.

Frigg, R. & Votsis, I. (2011). Everything You Always Wanted to Know about Structural Realism but were Afraid to Ask. *European Journal for Philosophy of Science*, 1(2), 227–276. https://doi.org/10.1007/s13194-011-0025-7.

Frisch, M. (2015). Predictivism and Old Evidence: A Critical Look at Climate Model Tuning. *European Journal for the Philosophy of Science*, 5(2), 171–190. https://doi.org/10.1007/s13194-015-0110-4.

Frost-Arnold, G. (2010). The No-Miracles Argument for Realism: Inference to an Unacceptable Explanation. *Philosophy of Science*, 77(1), 35–58. https://doi.org/10.1086/650207.

Geller, R.J., Jackson, D.D., Kagan, Y.Y. & Mulargia, F. (1997). Earthquakes Cannot Be Predicted. *Science*, 275(5306), 1616. https://doi.org/10.1126/science.275.5306.1616.

Glymour, C. (1980). *Theory and Evidence*. Princeton: Princeton University Press.

Glymour, C. (2008). Review: The Paradox of Predictivism. *Notre Dame Philosophical Reviews*.

Gross, C. (2016). Scientific misconduct. *Annual Review of Psychology*, 67, 693–711. https://doi.org/10.1146/annurev-psych-122414-033437.

Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814563.

Hájek, A., & Joyce, J. M. (2008). Confirmation. In S. Psillos & M. Curd (eds.), *The Routledge companion to the philosophy of science*. London: Routledge, 115–128.

Harker, D. (2006). Accommodation and Prediction: The Case of the Persistent Head. *The British Journal for the Philosophy of Science*, 57(2), 309–321. https://doi.org/10.1093/bjps/axl004.

Harker, D. (2008). On the Predilections for Predictions. *The British Journal for the Philosophy of Science*, 59(3), 429–453. https://www.jstor.org/stable/40072294.

Harker, D. (2011). Review: The Paradox of Predictivism. *The British Journal for the Philosophy of Science*, 62(1), 219–223. https://doi.org/10.1093/bjps/axq027.

Harker, D. (2013). How to Split a Theory: Defending Selective Realism and Convergence without Proximity. *The British Journal for the Philosophy of Science*, 64, 79–106. https://doi.org/10.1093/bjps/axr059.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T. & Jennions, M.D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, 13(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Heirtzler, J.R. & Le Pichon, X. (1974). FAMOUS: A Plate Tectonics Study of the Genesis of the Lithosphere. *Geology*, 2(6), 273–274. https://doi.org/10.1130/0091-7613(1974)2<273:FAPTSO>2.0.CO;2.

Henderson, L. (2017). The no miracles argument and the base rate fallacy. *Synthese*, 194, 1295–1302. https://doi.org/10.1007/s11229-015-0995-7.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press.

Hess, H.H. (1962). History of ocean basins. In A.E.J. Engel, H.L. James & B.F. Leonard (eds.), *Petrologic Studies: A Volume in Honor of A. F. Buddington*. New York: The Geological Society of America, 599–620. https://doi.org/10.1130/Petrologic.1962.

Hitchcock, C. & Sober, E. (2004). Prediction Versus Accommodation and the Risk of Overfitting. *British Journal for the Philosophy of Science*, 55(1), 1–34. https://doi.org/10.1093/bjps/55.1.1.

Hollenbeck, J.R. & Wright, P.M. (2017). Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data. *Journal of Management*, 43(1), 5–18. https://doi.org/10.1177/0149206316679487.

Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge University Press.

Howson, C. (1988). Accommodation, Prediction and Bayesian Confirmation Theory. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1988*, 2, 381–392. https://doi.org/10.1086/psaprocbienmeetp.1988.2.192899.

Howson, C. (1990). Fitting Your Theory to the Facts: Probably Not Such a Bad Thing After All. In C.W. Savage (ed.), *Scientific Theories. Minnesota Studies in the Philosophy of Science. Vol. XIV*. Minneapolis: University of Minnesota Press, 224–244.

Howson, C. (2000). *Hume's Problem: Induction and the Justification of Belief*. Oxford: Clarendon Press. https://doi.org/10.1093/0198250371.001.0001.

Howson, C. (2013). Exhuming the No-Miracles Argument. *Analysis*, 73(2), 205–211. https://doi.org/10.1093/analys/ant012.

Howson, C. & Franklin, A. (1991). Maher, Mendeleev and Bayesianism. *Philosophy of Science*, 58(4), 574–585. https://doi.org/10.1086/289641.

Howson, C. & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach. 2nd edition*. Chicago: Open Court.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2.

Hunt, J.C. (2012). On Ad Hoc Hypotheses. *Philosophy of Science*, 79(1), 1–14. https://doi.org/10.1086/663238.

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124.

John, L.K., Loewenstein, K. & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532. https://doi.org/10.1177/0956797611430953.

Kahn, J.A., Landsburg, S.E. & Stockman, A.C. (1992). On Novel Confirmation. *The British Journal for the Philosophy of Science*, 43(4), 503–516. https://doi.org/10.1093/bjps/43.4.503.

Keas, M.N. (2018). Systematizing the theoretical virtues. *Synthese*, 195, 2761–2793. https://doi.org/10.1007/s11229-017-1355-6.

Kerr, N.L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4.

Keynes, J.M. (1921). *A Treatise on Probability*. London: Macmillan.

Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.

Ladyman, J. (1999). Review: A Novel Defense of Scientific Realism. *The British Journal for the Philosophy of Science*, 50(1), 181–188. https://www.jstor.org/stable/40072217.

Lakatos, I. (1968). Changes in the Problem of Inductive Logic. In I. Lakatos (ed.), *The Problem of Inductive Logic*. Amsterdam: North-Holland Publishing Company, 315–417.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (eds.), *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Cambridge: Cambridge University Press, 91–195. https://doi.org/10.1017/CBO9781139171434.009.

Lange, M. (2001). The Apparent Superiority of Prediction to Accommodation: a Reply to Maher. *British Journal for the Philosophy of Science*, 52(3), 575–588. https://doi.org/10.1093/bjps/52.3.575.

Laudan, L. (1981). A Confutation of Convergent Realism. *Philosophy of Science*, 48(1), 19–49. https://doi.org/10.1086/288975.

Laudan, L. (1990). Demystifying Underdetermination. C.W. Savage (ed.), *Scientific Theories. Minnesota Studies in the Philosophy of Science, vol. 14*. Minneapolis: University of Minnesota Press, 267–297.

Laudan, L. & Leplin, J. (1991). Empirical Equivalence and Underdetermination. *Journal of Philosophy*, 88, 449–472. https://doi.org/10.2307/2026601.

Lee, W-Y. (2013). Akaike's Theorem and weak predictivism in science. *Studies in History and Philosophy of Science*, 44, 594–599. http://dx.doi.org/10.1016/j.shpsa.2013.06.001.

Lente, G. (2019). Where Mendeleev was wrong: predicted elements that have never been found. *ChemTexts*, 5(17). https://doi.org/10.1007/s40828-019-0092-5.

Leplin, J. (1997). *A Novel Defense of Scientific Realism*. Oxford: Oxford University Press.

Leplin, J. (2009). Review: The Paradox of Predictivism. *The Review of Metaphysics*, 63(2), 455–457.

Lewis, P.J. (2001). Why the Pessimistic Induction is a Fallacy, *Synthese*, 129(3), 371–380. https://doi.org/10.1023/A:1013139410613.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.W., Villringer, A. & Margulies, D.S. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, 148, 179–188. https://doi.org/10.1016/j.neuroimage.2016.11.005.

Lipton, P. (2004). *Inference to the best explanation. 2nd edition*. London/New York: Routledge.

Lyons, T.D. (2002). Scientific Realism and the Pessimistic Meta-Modus Tollens. In S. Clarke & T.D. Lyons (eds.), *Recent Themes in the Philosophy of Science. Australasian Studies in History and Philosophy of Science, vol 17*. Dordrecht: Springer, 63–90. https://doi.org/10.1007/978-94-017-2862-1_4.

Lyons, T.D. (2006). Scientific Realism and the Stratagema de Divide et Impera. *The British Journal for the Philosophy of Science*, 57(3), 537–560. https://doi.org/10.1093/bjps/axl021.

Lyons, T.D. (2017). Epistemic selectivity, historical threats, and the non-epistemic tenets of scientific realism. *Synthese*, 194, 3203–3219. https://doi.org/10.1007/s11229-016-1103-3.

Magnus, P.D. & Callender, C. (2004). Realist Ennui and the Base Rate Fallacy. *Philosophy of Science*, 71(3), 320–338. https://doi.org/10.1086/421536.

Maher, P. (1988). Prediction, Accommodation, and the Logic of Discovery. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1988,* 1, 273–285. https://doi.org/10.1086/psaprocbienmeetp.1988.1.192994.

Maher, P. (1990). How Prediction Enhances Confirmation. In J.M Dunn & A. Gupta (eds.), *Truth or Consequences: Essays in Honor of Nuel Belnap*. Dordrecht: Kluwer, 327–343. https://doi.org/10.1007/978-94-009-0681-5.

Maher, P. (1993). Howson and Franklin on Prediction. *Philosophy of Science*, 60(2), 329–340. https://doi.org/10.1086/289736.

Mayo, D.G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago/London: University of Chicago Press.

Mayo, D.G. (2009). An Ad Hoc Save of a Theory of Adhocness? Exchanges with John Worrall. In D.G. Mayo & A. Spanos (eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge: Cambridge University Press, 155–169. https://doi.org/10.1017/CBO9780511657528.006.

Mayo, D.G. (2014). Some surprising facts about (the problem of) surprising facts (from the Dusseldorf Conference, February 2011). *Studies in History and Philosophy of Science Part A*, 45, 79–86. https://doi.org/10.1016/j.shpsa.2013.10.005.

McCain, K. & Poston, T. (2014). Why Explanatoriness Is Evidentially Relevant. *Thought: A Journal of Philosophy*, 3(2), 145–153.

Mill, J.S. (1843). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, 2 vols*. London: John W. Parker.

Mizrahi, M. (2015). Historical Inductions: New Cherries, Same Old Cherry-picking. *International Studies in the Philosophy of Science*, 29(2), 129–148. https://doi.org/10.1080/02698595.2015.1119413.

Motyl, M., Demos, A.P., Carsel, T.S., Hanson, B E., Melton, Z.J., Mueller, A.B., Prims, J.P., Sun, J., Washburn, A.N., Wong, K.M., Yantis, C. & Skitka, L.J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, 113(1), 34–58. http://dx.doi.org/10.1037/pspa0000084.

Murphy, K.R. & Aguinis, H. (2019). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*, 34, 1–17. https://doi.org/10.1007/s10869-017-9524-7.

Musgrave, A. (1974). Logical Versus Historical Theories of Confirmation. *British Journal for the Philosophy of Science*, 25(1), 1–23. https://doi.org/10.1093/bjps/25.1.1.

Musgrave, A. (1988). The Ultimate Argument for Scientific Realism. In R. Nola (eds.), *Relativism and Realism in Science. Australasian Studies in History and Philosophy of Science, vol 6*. Dordrecht: Springer, 229–252. https://doi.org/10.1007/978-94-009-2877-0_10.

Muthukrishna, M. & Henrich, J. (2019). A problem in theory. *Nature Human Behavior*, 3, 221–229. https://doi.org/10.1038/s41562-018-0522-1.

Ne'eman, Y. & Kirsch, Y. (1986). *The Particle Hunters*. Cambridge: Cambridge University Press.

Niiniluoto, I. (2017). Optimistic realism about scientific progress. *Synthese*, 194, 3291–3309. https://doi.org/10.1007/s11229-015-0974-z.

Niiniluoto, I. (2019). Scientific Progress. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition*). URL = <https://plato.stanford.edu/archives/win2019/entries/scientific-progress/>.

Northcott, R. (2017). When are Purely Predictive Models Best? *Disputatio*, 9(47), 631–656. https://doi.org/10.1515/disp-2017-0021.

Nosek, B.A., Ebersole, C.R., DeHaven, C.A. & Mellor, D.T. (2018). The Preregistration Revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716.

Partington, J.R. & McKie, D. (1938). Historical Studies on the Phlogiston Theory. II. The Negative Weight of Phlogiston. *Annals of Science*, 3(1), 1–58. https://doi.org/10.1080/00033793800200781.

Peirce, C.S. (1883). A theory of probable inference. In C.S. Peirce (ed.), *Studies in Logic, By Members of the Johns Hopkins University*. Boston: Little, Brown & Company, 126–181. https://doi.org/10.1037/12811-007.

Peters, D. (2014). What Elements of Successful Scientific Theories Are the Correct Targets for 'Selective' Scientific Realism? *Philosophy of Science*, 81(3), 377–397. https://doi.org/10.1086/676537.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York and Evanston: Harper and Row.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge. https://doi.org/10.4324/9780203979648.

Psillos, S. (2000). The Present State of the Scientific Realism Debate. *The British Journal for the Philosophy of Science*, 51, 705–728. https://www.jstor.org/stable/3541614.

Psillos S. (2006). Thinking About the Ultimate Argument for Realism. In C. Cheyne & J. Worrall (eds.), *Rationality and Reality: Conversations with*

Alan Musgrave. Dordrecht: Springer, 133–156. https://doi.org/10.1007/1-4020-4207-8_8.

Psillos, S. (2011). The Scope and Limits of the No Miracles Argument. In D. Dieks, W.J. Gonzalez, S. Hartmann, T. Uebel & M. Weber (eds.), *Explanation, Prediction, and Confirmation*. Dordrecht: Springer, 23–35. https://doi.org/10.1007/978-94-007-1180-8.

Psillos, S. (2020). Resisting scientific anti-realism. Review of "Resisting Scientific Realism," by K. Brad Wray. *Metascience*, 29, 17–24. https://doi.org/10.1007/s11016-020-00488-1.

Putnam, H. (1975). *Mathematics, Matter and Method*. Cambridge: Cambridge University Press.

Quine, W.V.O. (1951). Two Dogmas of Empiricism. *The Philosophical Review*. 60(1), 20-43. https://doi.org/10.2307/2181906.

Roche, W. & Sober, E. (2013). Explanatoriness is evidentially irrelevant, or inference to the best explanation meets Bayesian confirmation theory. *Analysis*, 73(4), 659–668. https://doi.org/10.1093/analys/ant079.

Roche, W. & Sober, E. (2014). Explanatoriness and Evidence: A Reply to McCain and Poston. *Thought: A Journal of Philosophy*, 3, 193–199. https://doi.org/10.1002/tht3.128.

Roche, W. & Sober, E. (2017) Is Explanatoriness a Guide to Confirmation? A Reply to Climenhaga. *Journal for the General Philosophy of Science*, 48, 581–590. https://doi.org/10.1007/s10838-016-9357-5.

Rossetter, T. (2018). Realism on the rocks: Novel success and James Hutton's theory of the earth. *Studies in History and Philosophy of Science Part A*, 67, 1–13. https://doi.org/10.1016/j.shpsa.2017.10.005.

Rowbottom, D.P. (2019a). *The Instrument of Science: Scientific Anti-realism Revitalised*. Abingdon: Routledge.

Rowbottom, D.P. (2019b). Scientific realism: what it is, the contemporary debate, and new directions. *Synthese*, 196, 451–484. https://doi.org/10.1007/s11229-017-1484-y.

Rowley, D.B., Forte, A.M., Rowan, C.J., Glišović, P., Moucha, R., Grand, S.P. & Simmons, N.A. (2016). Kinematics and dynamics of the East Pacific Rise linked to a stable, deep-mantle upwelling. *Science advances*, 2(12), e1601107. https://doi.org/10.1126/sciadv.1601107.

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology*, 21(4), 308–320. https://doi.org/10.1037/gpr0000128.

Rubin, M. (forthcoming). The Costs of HARKing. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz050.

Saatsi, J. (2005a). Reconsidering the Fresnel-Maxwell Case Study. *Studies in the History and Philosophy of Science*, 36(3), 509–38. https://doi.org/10.1016/j.shpsa.2005.07.007.

Saatsi, J. (2005b). On the Pessimistic Induction and Two Fallacies. *Philosophy of Science*, 72(5), 1088–1098. https://doi.org/10.1086/508959.

Saatsi, J. (2017). Replacing recipe realism. *Synthese*, 194, 3233-3244. https://doi.org/10.1007/s11229-015-0962-3.

Saatsi, J. (2018). Realism and the limits of explanatory reasoning. In J. Saatsi (ed.), *The Routledge Handbook of Scientific Realism*. Abingdon: Routledge, 200-212.

Scerri, E.R. & Worrall, J. (2001). Prediction and the periodic table. *Studies in History and Philosophy of Science Part A*, 32(3), 407–452. https://doi.org/10.1016/S0039-3681(01)00023-1.

Scerri, E.R. (2007). *The Periodic Table: Its Story and its Significance*. New York: Oxford University Press.

Scerri, E.R. (2016). *A Tale of Seven Scientists and a New Philosophy of Science*. New York: Oxford University Press.

Schindler, S. (2008). Use Novel Predictions and Mendeleev's Periodic Table: Response to Scerri and Worrall (2001). *Studies in the History and Philosophy of Science Part A*, 39(2), 265–269. https://doi.org/10.1016/j.shpsa.2008.03.008.

Schindler, S. (2014). Novelty, coherence, and Mendeleev's periodic table. *Studies in History and Philosophy of Science Part A*, 45, 62–69. http://dx.doi.org/10.1016/j.shpsa.2013.10.007.

Schippers, M & Schurz, G. (2020). Genuine Confirmation and Tacking by Conjunction. *British Journal for the Philosophy of Science*, 71(1), 321–352. https://doi.org/10.1093/bjps/axy005.

Schlesinger, G.N. (1987). Accommodation and Prediction. *Australasian Journal of Philosophy*, 65(1), 33–42. https://doi.org/10.1080/00048408712342751.

Schurz, G. (1991). Relevant deduction. *Erkenntnis*, 35, 391–437. https://doi.org/10.1007/BF00388295.

Schurz, G. (2014). Bayesian Pseudo-Confirmation, Use-Novelty, and Genuine Confirmation. *Studies in History and Philosophy of Science Part A*, 45, 87–96. https://doi.org/10.1016/j.shpsa.2013.10.008.

Schurz, G. & Weingartner, P. (2010). Zwart and Franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude. *Synthese*, 172, 415–436. https://doi.org/10.1007/s11229-008-9399-2.

Shaw, J.D. (2017). Advantages of Starting with Theory. *Academy of Management Journal*, 60(3), 819–822. https://doi.org/10.5465/amj.2017.4003.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Spottiswode, W. (1883). Presidential Address. *Proceedings of the Royal Society*, 34, 303–329.

Sprenger, J. (2013). The role of Bayesian philosophy within Bayesian model selection. *European Journal for the Philosophy of Science*, 3, 101–114. https://doi.org/10.1007/s13194-012-0059-5.

Stainforth, D.A., Allen, M.R., Tredger, E.R. & Smith, LA. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A*, 365, 2145–2161. https://doi.org/10.1098/rsta.2007.2074.

Stanford, P.K. (2006). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press. https://doi.org/10.1093/0195174089.001.0001.

Stanford, P.K. (2009). Author's Response. *Metascience*, 18(3), 379–90.

Stanford, P.K. (2020). Resisting Scientific Realism with or Without van Fraassen's Darwinian Explanation. *Metascience*, 29(1), 25–31. https://doi.org/10.1007/s11016-020-00489-0.

Steele, K. & Werndl, C. (2013). Climate Models, Calibration, and Confirmation. *The British Journal for the Philosophy of Science*, 64(3), 609–635. https://doi.org/10.1093/bjps/axs036.

Steele, K. & Werndl, C. (2016). The diversity of model tuning practices in climate science. *Philosophy of Science*, 83(5), 1133–1144. https://doi.org/10.1086/687944.

Steele, K. & Werndl, C. (2018). Model-Selection Theory: The Need for a More Nuanced Picture of Use-Novelty and Double-Counting. *The British Journal for the Philosophy of Science*, 69(2), 351–375. https://doi.org/10.1093/bjps/axw024.

Stewart, P.J. (2019). Mendeleev's Predictions: Success and Failure. *Foundations of Chemistry*, 21(1), 3–9. https://doi.org/10.1007/s10698-018-9312-0.

Stroebe, W., Postmes, T. & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7, 670–688. https://doi.org/10.1177/1745691612460687.

The New York Times. (1919). Lights All Askew In The Heavens. https://timesmachine.nytimes.com/timesmachine/1919/11/10/118180487.html?pageNumber=17.

Tulodziecki, D. (2017). Against Selective Realism(s). *Philosophy of Science*, 84(5), 996–1007. https://doi.org/10.1086/694004.

Ulrich, R. & Miller, J. (2020). Meta-Research: Questionable research practices may have little effect on replicability. *eLife*, 9, e58237. https://doi.org/10.7554/eLife.58237.

Vancouver, J.B. (2018). In Defense of HARKing. *Industrial and Organizational Psychology*, 11(1), 73–80. https://doi.org/10.1017/iop.2017.89.

van Fraassen, B.C. (1980). *The Scientific Image*. Oxford: Oxford University Press. https://doi.org/10.1093/0198244274.001.0001.

Vickers, P. (2013). A Confrontation of Convergent Realism. *Philosophy of Science*, 80(2), 189–211. https://doi.org/10.1086/670297.

Vickers, P. (2017). Understanding the selective realist defense against the PMI. *Synthese*, 194, 3221–3232. https://doi.org/10.1007/s11229-016-1082-4.

Vickers, P. (2018). Quo Vadis Selective Scientific Realism? *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 9(1), 118–121. https://doi.org/10.4245/sponge.v9i1.28056.

Vickers, P. (2019). Towards a realistic success-to-truth inference for scientific realism. *Synthese*, 196, 571–585. https://doi.org/10.1007/s11229-016-1150-9.

Vine, F.J. & Matthews, D.H. (1963). Magnetic anomalies over the oceanic ridges. *Nature*, 199, 947–949. https://doi.org/10.1038/199947a0.

Votsis, I. (2014). Objectivity in Confirmation: Post Hoc Monsters and Novel Predictions. Studies in History and Philosophy of Science Part A, 45, 70–78. https://doi.org/10.1016/j.shpsa.2013.10.009.

Watts, D.J., Beck, E.D., Bienenstock, E.J., Bowers, J., Frank, A., Grubesic, A., Hofman, J.M., Rohrer, J. & Salganik, M. (2018). Explanation, Prediction, and Causality: Three Sides of the Same Coin? *OSF Preprints*. https://doi.org/10.31219/osf.io/u6vz5.

Weisberg, M. (2007). Three Kinds of Idealization. *The Journal of Philosophy*, 104(12), 639–659.

Whewell, W. (1840). *The Philosophy of the Inductive Sciences, Founded Upon Their History, 2 vols*. London: John W. Parker.

White, R. (2003). The Epistemic Advantage of Prediction over Accommodation. *Mind*, 112(448), 653–683. https://doi.org/10.1093/mind/112.448.653.

Williamson, J. (2010). *In defense of objective Bayesianism*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199228003.001.0001.

Wilson, E.B. (1952). *An Introduction to Scientific Research*. New York: McGraw-Hill.

Wilson, B.M. & Wixted, J.T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, 1, 186–197. https://doi.org/10.1177/2515245918767122.

Winsberg, E.B. (2018). *Philosophy and climate science*. New York: Cambridge University Press. https://doi.org/10.1017/9781108164290.

Worrall, J. (1985). Scientific Discovery and Theory-Confirmation. In J.C. Pitt (ed.), *Change and Progress in Modern Science*. Dordrecht: Reidel, 301–331. https://doi.org/10.1007/978-94-009-6525-6.

Worrall, J. (1989a). Structural Realism: The Best of Both Worlds? *Dialectica*, 43, 99–124. https://www.jstor.org/stable/42970613.

Worrall, J. (1989b). Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories. In D. Gooding, T. Pinch & S. Schaffer (eds.), *The Uses of Experiment: Studies in the Natural Sciences*. Cambridge: Cambridge University Press, 135–157.

Worrall, J. (2002). New Evidence for Old. In P. Gärdenfors, J. Wolenski & K. Kijania-Placek (eds.), *In the Scope of Logic, Methodology and Philosophy of Science: Volume One of the 11th International Congress of Logic, Methodology and Philosophy of Science, Cracow, August 1999*. Dordrecht: Kluwer Academic Publishers, 191–209.

Worrall, J. (2005). Miracles, Pessimism and Scientific Realism. Unpublished manuscript. URL = < https://philpapers.org/archive/WORMPA.pdf>.

Worrall, J. (2006). Theory-Confirmation and History. In C. Cheyne & J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*. Dordrecht: Springer, 31–61. https://doi.org/10.1007/1-4020-4207-8.

Worrall, J. (2009). Error, Tests, and Theory Confirmation. In D.G. Mayo & A. Spanos (eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge: Cambridge University Press, 125–154. https://doi.org/10.1017/CBO9780511657528.006.

Worrall, J. (2011). The No Miracles Intuition and the No Miracles Argument. In D. Dieks, W.J. Gonzalez, S. Hartmann, T. Uebel & M. Weber (eds.), *Explanation, Prediction, and Confirmation*. Dordrecht: Springer, 11–21. https://doi.org/10.1007/978-94-007-1180-8.

Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science Part A*, 45, 54–61. https://doi.org/10.1016/j.shpsa.2013.10.001.

Woznyj, H.M., Grenier, K., Ross, R., Banks, G.C. & Rogelberg, S.G. (2018). Results-blind review: a masked crusader for science. *European Journal for Work and Organizational Psychology*, 27(5), 561–576. https://doi.org/10.1080/1359432X.2018.1496081.

Wray, K.B. (2013). The pessimistic induction and the exponential growth of science reassessed. *Synthese*, 190(18), 4321–4330. https://doi.org/10.1007/s11229-013-0276-2.

Wray, K.B. (2015). Pessimistic Inductions: Four Varieties. *International Studies in the Philosophy of Science*, 29(1), 61–73. http://dx.doi.org/10.1080/02698595.2015.1071551.

Wray, K.B. (2018). *Resisting Scientific Realism*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108231633.

Wray, K.B. (2019). What to make of Mendeleev's predictions? *Foundations of Chemistry*, 21(2), 139–143. https://doi.org/10.1007/s10698-018-9313-z.

Yarkoni, T. & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393.

Ylikoski, P. & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, 148, 201–219. https://doi.org/10.1007/s11098-008-9324-z.

Zahar, E. (1973). Why did Einstein's Programme supersede Lorentz's? (I). *British Journal for the Philosophy of Science*, 24(2), 95–123. https://doi.org/10.1093/bjps/24.2.95.