# How to be a scientific realist (if at all):
## A study of partial realism

Dean Peters

A thesis submitted to the Department of Philosophy, Logic and
Scientific Method of the London School of Economics for the
degree of Doctor of Philosophy, London, October 2012

## Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own. The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of **86,011** words, including footnotes but excluding references.

Dean Peters

## Abstract

"Partial realism" is a common position in the contemporary philosophy of science literature. It states that the "essential" elements of empirically successful scientific theories accurately represent corresponding features the world. This thesis makes several novel contributions related to this position. Firstly, it offers a new definition of the concept of “empirical success”, representing a principled merger between the use-novelty and unification accounts. Secondly, it provides a comparative critical analysis of various accounts of which elements are “essential” to the success of a theory, including structural realism and the *divide et impera* strategy. A novel account of essentialness, entitled the “empirically successful sub-theory account”, is defended. Thirdly, it is argued that the realism/anti-realism debate should put to the side metaphysical questions and focus instead on partial realism’s commitment to the continuity of science. Because this commitment lacks metaphysical implications, it is referred to as “deflationary realism”. Anti-realists cannot reject deflationary realism as a matter of a priori principle; its overall viability (and therefore that of partial realism) can therefore only be assessed by a careful examination of the history of science. Finally, another consequence of partial realism, named "partial rationalism", is defended. Partial rationalism states that, in cases where several competing theories have been suggested, scientists are rational just in case they accept the essential elements of each of the scientific theories on offer. This novel position subverts the existing literature on scientific “revolutions”, as it sometimes demands that scientists devise a synthesis between competing scientific theories, instead of “choosing” only one. The philosophical points defended in this thesis are illustrated and supported by case studies from the history of science, including Fresnel’s wave theory of light, the Copernican revolution, the “neo-Darwinian synthesis” in evolutionary biology, the “prion revolution” in molecular biology, the miasma theory of disease, and the chemical revolution

## Outline of thesis

## Table of Contents

## Acknowledgements

Many thanks are due to my supervisors, John Worrall and Roman Frigg. Despite the many deadlines I have missed, and my general propensity for leaving things to the very last minute, they have been extremely patient and supportive with me. They had confidence in me even when I didn't, and they pushed me when I needed it. They have been very forthcoming with their time and their philosophical expertise, right up until the last days of the writing-up process.

I am especially indebted to John. He encouraged me to apply for the MPhil/PhD when I was still in the early months of my MSc degree, my first real exposure to academic philosophy. Even then, he provided some valuable supervision for my MSc thesis. Anyone who reads this thesis will discern the intellectual debt it owes to John. While we have not always agreed, and still don't agree on several points, he has always been willing to discuss ideas with me. His straightforward criticism has led me to change my mind on many issues, and made my arguments stronger when I didn't

I have been financially supported throughout by the London School of Economics and Political Science (LSE), under the PhD Studentship scheme. This research would have been impossible without that support, and I am extremely grateful. I am also grateful to Richard Bradley, Tom Chivers and Armin Schulz, who very generously assisted me with the administration involved in this.

I thank all the teachers and students of philosophy at the LSE, and at the University of London generally. I have learned a huge amount from many of them, and even more of them have kindly tolerated my thinking out loud in public. Special thanks must go to the members of the Philosophy of Science reading group at the LSE, especially Seamus Bradley, Jonathan Everett and Steph Ratcliffe; and the Gender and Philosophy reading group at Kings College London, especially Mary Carman, Alexander Davies, Marion Godman, Elselijn Kingma and Kalbir Sohi. I

have had many valuable discussions in the all-London History and Philosophy of Science seminar series, held every summer at the LSE. Participants in this group have included Hasok Chang, Nancy Cartwright, Foad Dizadji-Bahmani and Miklos Redei. I was first exposed to Hasok's ideas about the phlogiston theory of chemistry in this group, and I was sufficiently intrigued that I have written a section of the following thesis on this topic. I have had very many stimulating interactions with permanent and visiting students at the Department of Philosophy, Logic and Scientific Method at the LSE. Of the people not already mentioned, I especially wish to thank Susanne Burri, Mareile Drechsler, Ben Ferguson, Silvia Milano, Alice Obrecht, Francesca Perro, Orri Steffansson and Mischa van den Brandhof. I am grateful to the 2011-2012 class of PH201: Philosophy of Science at the LSE. I was fortunate to teach this course while I was writing the thesis, and the level of engagement I enjoyed with the students was extremely helpful. I also wish to think David Papineau, who was always willing to make time for discussions with me during my MSc and the early years of the PhD. I learned a huge amount from him.

Many of the ideas in this thesis have arisen out of presentations I delivered at conferences. Interaction with the participants at each of these conferences often improved the corresponding papers enormously. I have also been fortunate enough to maintain continuing contact with many of them. The original idea of "partial realism" and the miasma theory of disease case study were presented at the British Society for the Philosophy of Science meeting in Dublin in July 2010 and at the British Society for the History of Science Postgraduate Conference in Manchester in January 2011. I gave a sketch of the idea that eventually developed into Chapter 2 at the Novel Predictions conference, which took place in Dusseldorf in February 2011. The comments of Deborah Mayo, Kit Patrick, Samuel Schindler and Gerhard Schurz were particularly insightful. Chapter 3 has been much improved by a lively exchange I have enjoyed with Pete Vickers. Pete has recently hosted two workshops on scientific realism at Durham University, in May and September 2012, and I have had extremely productive conversations with him and the other participants there, particularly Juha Saatsi. Much of Chapter 4 was written for the conference Beyond Rationality III - Resistance and the Practice of

Rationality, hosted at the LSE in November 2010. My thanks are due to Roman here, too, for asking me to speak at this conference in his place. My thanks also to Rom Harre, Martin Bauer and Carl Jensen for hosting this excellent cross-disciplinary conference. Finally, parts of Chapters 4 and 5 were written for the Progress of Science conference held in Tilburg in April 2012. I am especially grateful to Kirsten Walsh for her valuable feedback there.

None of this could have been accomplished without support from my friends and family. My thanks to Grant Theron for his endlessly (and just endless) amusing chatter, and for showing me the ins and outs of academia. Caeli Christianson supported me enormously, when she could. I will always remember it. Rony Hacohen has been an amazing source of support and good humour over the final stretch. The last few months have not been easy, and I am grateful to all my friends, old and new, who have been there with me in the thick of it. My biggest thanks go to my parents. I wish that my father, Douglas Peters, was alive to share this with me. My mother, Silvana Peters, is one of the most amazing people I know. She has always supported my education to the hilt, whatever the cost, and however far it took her only child away from her. I will always be grateful to her.

## **Thesis overview**

Scientific realism claims, roughly, that our best scientific theories are at least approximately true. In response to various criticisms of this naive formulation, most advocates of realism today advocate some or other version of what I call “partial realism” (PR). This states that only the "essential" elements of empirically successful scientific theories accurately represent corresponding features of the mind-independent world. In this thesis, I also articulate a particular consequence of PR, which I call “deflationary realism” (DR). This states that the essential elements of empirically successful theories will be preserved in successor theories, but offers no metaphysical explanation of why this is the case. DR is a purely empirical claim about the history of science, past and future.

My aims in this thesis are three-fold. Firstly, by synthesising arguments already present in the literature and devising new arguments, I will attempt to articulate as defensible an account of PR as possible. Secondly, I will propose DR both as an elegant vehicle for furthering the realism/anti-realism debate, and as a defensible position in its own right. Thirdly, as means of assessing DR’s claims about the history of science, I will examine several historical case studies.

In Chapter 1, it is shown that PR emerges as a reasonable response to the crucial arguments for and against scientific realism, respectively the “no-miracles argument” (NMA) and the “pessimistic meta-induction” (PMI). However, PR, as it is expressed in this chapter, is more a broad philosophical approach than a specific position. This is because different authors disagree on the appropriate interpretations of the terms “empirical success” and “essential”. Thus, the major project of Chapters 2 and 3 is to provide defensible interpretations of these terms.

The idea of empirical success is addressed in Chapter 2. One of the major typological differences between different accounts of empirical success is that between logical approaches, which emphasise the objective relations between theory and evidence, and historical approaches, which emphasise the context in which a particular piece of evidence for the theory is introduced. The currently dominant approach to empirical success in the scientific realism literature emphasises the value of novel predictive success, thus representing a type of historical approach. However, it is shown in this chapter that historical accounts are quite generally unsatisfactory, as the fact that a theory satisfies such an account is not sufficient to regard it as confirmed under the NMA. A general approach with the name "weak predictivism" is therefore advocated instead of novel prediction approaches. Weak predictivist accounts regard a theory as well-confirmed to the extent that it effectively unifies verified empirical results. Under these accounts, successful novel prediction is merely, under some circumstances, an *indicator* that a theory has unifying power. One specific weak predictivist account, named the "unification view" (UV), and based on Worrall's version of the use-novelty (UN) account of prediction is defended. The UV states that a theory is confirmed to the extent that it gives rise to more verified empirical results than are required to construct it.

The question of what parts of a theory are essential to the empirical success of that theory is addressed in Chapter 3. The bulk of this chapter consists of a systematic critical analysis of particular accounts of essentialness that have been proposed in the literature. The NMA, if valid, licences the inference from some empirical success to the truth of the propositions "responsible" for that success. Since the motivating force behind PR is the NMA, the most important principle in comparing various accounts of essentialness is the following: whatever elements are picked out by the favoured account must be sufficient to *explain* the empirical success of the theory. This account, moreover, should not exclude any theoretical propositions which would, if true, genuinely contribute to an explanation. On the basis of these two desiderata, direct reference theories and structural realism are rejected as being too "inclusive", whereas entity realism, phenomenological realism and the working posits idea (under the most reasonable interpretation so far offered) are rejected as being too "exclusive". I go on to defend an account of essentialness that is based on the account of empirical success endorsed in Chapter 2, and which is therefore named the "empirically successful sub-theory account" (ESSA).

The ESSA characterises as essential just those theoretical posits which give rise to more lower-level posits than required to construct them.

The general form of DR, like that of PR, is underspecified because it contains the vague terms “empirical success” and “essential”. Despite the fact that DR lacks the metaphysical commitments which give rise to the interpretations of these terms proposed in Chapters 2 and 3, it is nevertheless suggested that DR should adopt these interpretations. This ensures that any dispute over the accuracy of DR has relevance for the larger realist debate since, if DR is a logical consequence of PR, any refutation of the former is also a refutation of the latter. It is demonstrated in Chapter 1, however, that it is also possible to construct an “optimistic induction” for DR in its own right. If theoretical continuity (of the specified form) is thus established as a genuine regularity in the history of science, then scientists have a good pragmatic reason to theorise conservatively (in a certain way). Indeed, even the constructive empiricist might by this route come to have a reasoned commitment to the continuity of science.

The larger part of Chapters 4 and 5 is concerned with examining the tenability of DR (and thus PR) in light of various episodes from the history of science. In Chapter 4, the issue of continuity in science is connected explicitly to the related debate around rationality in cases of theory choice. In this chapter, two broad approaches will be compared. The first of these is “Kuhnian history” which states that, while objective epistemic standards constrain the choices of scientists, they do not completely determine them. The second approach is “rationalism”, which states that objective factors are, at least much of the time, sufficient to fully determine the rational course of action in cases of theory choice. A particular form of rationalism, called “partial rationalism”, is derivable from PR. It states that, during episodes of theory change, scientists are rational to accept those *parts* of theories responsible for empirical successes (and have largely done so). In consequence, this account states that it is sometimes rational to accept some *synthesis* of several competing theories, rather than choose between them. The claims of partial rationalism, it is argued, are largely borne out in the three case studies examined in this chapter, namely the Copernican revolution, the “neo-Darwinian” synthesis and the “prion revolution”.

In Chapter 5, the focus is more explicitly on potential counterexamples to DR, and thus to scientific realism generally. The chapter begins with a reasonably comprehensive list of the putative counterexamples suggested so far in the literature. From this list, two particular theories are selected for closer study, namely the miasma theory of disease and the phlogiston theory of chemistry. Given the focus in this thesis placed on *continuity* in science, these theories are examined in the contexts in which they were rejected by the scientific community in favour of newer theories. It is shown that, although there is no substantial continuity between the miasma theory and later theories of disease, this is not a challenge to the realist because the miasma theory was not substantively empirically successful. The phlogiston theory *was* empirically successful, and concepts that are structurally similar to those of phlogiston theory are found in modern electrochemistry. Analogues of these concepts are not, however, found in the immediate successor theory to phlogiston, namely Lavoisier’s oxygen theory. In consequence, it is suggested that DR is not always descriptively accurate in the short run, but may nevertheless be valuable as a source of methodological guidance.

# Chapter 1. Naive realism, partial realism and deflationary realism

## 1. Chapter overview

Scientific realism claims, roughly, that our best scientific theories are at least approximately true. The first few sections of this chapter introduce scientific realism in more detail, and state some of the key arguments for and against it. The major argument in support of scientific realism, called the no-miracles argument (NMA), is introduced in section 2, and some of the objections to it are discussed in section 3. The primary argument *against* scientific realism, the pessimistic meta-induction (PMI), and responses to it are discussed in section 4.

The major response of scientific realists to the PMI has been to adopt various, more moderate, forms of realism, which I group under the heading of "partial realism". Partial realism claims that we are only justified in regarding as true those *parts* of theories "essential" for their empirical success. I articulate a particular consequence of partial realism, which I call "deflationary realism" (DR). This states that those parts of theories essential for their empirical success will be preserved in successor theories. It is thus a commitment to the continuity of science. Both of these positions are outlined in section 5. The idea of a theoretical element being preserved across theory change is described in considerably more detail in section 6.

In defining a metaphysically deflationary view, the intent is not to deny that the metaphysical commitments of (partial) scientific realism are important. Indeed, these metaphysical comments also define the particular version of DR that will be examined in the later chapters of this thesis. The intent is to define a minimal position that can be tested simply as an empirical hypothesis. This minimal position remains relevant to the broader debate precisely because it is a logical consequence of metaphysically inflationary realism, and so a refutation of it counts as a refutation of the latter also.

Despite the emphasis on refutation, in section 7, a positive argumentative strategy in support of DR is proposed. Since DR is an empirical hypothesis, this has the form of an “optimistic induction”. Relatedly, in section 8, it is demonstrated that DR is compatible with constructive empiricism and how the constructive empiricist might therefore be led to adopt a stronger stance on the continuity of science.

Section 9 investigates an argumentative strategy whereby DR might be ‘re-inflated’ to yield a metaphysically substantial form of realism. This strategy starts from the supposition that the fact of theoretical continuity itself demands explanation, and then suggests that the best explanation of this fact is the truth of the theories concerned. This is named the “argument from continuity to truth”. Interestingly, the most difficult problem for this argument also represents a problem for the optimistic induction in support of DR itself. In consequence, some qualifications to the optimistic induction are suggested. The chapter is summarised in section 10.

## 2. Scientific realism and the no-miracles argument

Since much of the argumentation in this thesis will centre on the tenability of scientific realism, it is worth defining the commitments of this view carefully. An early articulation of a realist view is found in Poincaré:

> “[T]hese equations [of Fresnel’s theory of light] express relations, and if the equations remain true, it is because the relations preserve their reality. They teach us now, as they did then, that there is such and such a relation between this thing and that...” (Poincaré, 1902/1952, p. 161)

A similar view is expressed by Poincaré’s contemporary, Duhem (1906/1954). More recently, the anti-realist Van Fraassen defines scientific realism as the view that:

> “Science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true” (van Fraassen, 1980, p. 8).

And Chakravartty, in a survey article in the *Stanford Encyclopedia*, states that

> “Scientific realism is a [position that endorses belief in the reality of] ... whatever is described by our best scientific theories” (Chakravartty, 2011)

What is common to all these definitions is that, according to the scientific realist, the entities and processes described by our “best” or most “mature” scientific theories are *real*. Another way of putting this is that the statements of these theories are *true*.

The assertion that something is real is rendered a great deal more informative when conveyed alongside some idea of what the world would be like if this thing were *not* real. Scientific realism is often contrasted with instrumentalist or empiricist views. These assert, roughly, that our best scientific theories should not be thought of as accurate *representations* of the world, but simply as *tools* for dealing with it effectively. Not only is it unsound to infer that our best theories are true, but is a mistake to think of them as even *aiming* at the truth.

The dispute between scientific realists and anti-realists is not new. Indeed, both positions are represented in Copernicus’ great work, *De revolutionibus*, published in 1543. This work introduced the heliocentric model of the universe, and is one of the seminal works of the European Scientific Revolution. A Lutheran preacher named Osiander, who supervised the publication of this book, wrote a preface to it entitled “*To the Reader Concerning the Hypotheses of this Work”*. In this preface he states that:

> “[I]t is the duty of an astronomer to compose the history of the celestial motions through careful and expert study. Then he must conceive and

> devise the causes of these motions or hypotheses about them. Since he cannot in any way attain to the true causes, he will adopt whatever suppositions enable the motions to be computed correctly from the principles of geometry for the future as well as for the past. The present author has performed both these duties excellently. For these hypotheses need not be true nor even probable. On the contrary, if they provide a calculus consistent with the observations, that alone is enough." (Osiander in Copernicus, 1543/1978)

It is accepted by both Copernicus and Osiander that the heliocentric model proposed in this book was powerful and insightful. However, while there is every indication that Copernicus accepted it as a true description of the state of the heavens (although he does not state this explicitly), Osiander applies an instrumentalist, anti-realist interpretation.

Much more recently, and partly in response to the arguments of sophisticated anti-realists like van Fraassen, proponents of scientific realism have distinguished three separate commitments that together comprise realism. This formulation is suggested, for example, by Chakravartty (2011), Psillos (1999) and Niiniluoto, (1999). The three commitments are:

1. A metaphysical claim, that there is a mind-independent world. This rules out various idealist positions, but leaves many other issues quite open. Some authors argue that the world is characterised by a "natural kind" or "causal" structure, but other formulations are also compatible with the basic claim.
2. A semantic claim, that scientific theories are to be interpreted literally, and so can be true or false as descriptions of the mind-independent world.
3. An epistemic claim, that our 'most successful' scientific theories are at least approximately and/or partially true as descriptions of the world.

Notice that these claims are not logically independent – the second claim presupposes the first, and the third presupposes the first and second. In the existing literature, the third, epistemic, claim has in recent years been the focus of debate. This is partly because the most currently influential strain of anti-realism, namely van Fraassen's (1980) "constructive empiricism", accepts the second claim, is neutral on the first, but denies the third. Moreover, as we shall see below, some *prima facie* very powerful arguments have been levelled both for and against the third claim, and much effort has consequently been directed towards assessing the soundness of these arguments. For the purposes of this thesis therefore, I will follow the broad trend in the literature and assume that the metaphysical and semantic claims of scientific realism are granted, focusing narrowly on the epistemic claim.

The most intuitively powerful argument in favour of the epistemic claim of scientific realism is the so-called "no-miracles argument" (NMA). The most frequently-cited example of this argument, and that which provides its name, is given by Putnam:

> "The positive argument for realism is that it is the only philosophy that does not make the success of science a miracle. That terms in mature scientific theories typically refer ..., that the theories accepted in a mature science are typically approximately true, that the same terms can refer to the same even when they occurs in different theories—these statements are viewed not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate description of science and its relations to its objects" (Putnam 1975a, p. 73).

We should, in other words, believe in the literal truth of most successful scientific theories precisely because their truth would *explain* this success. It is worth noting that several proponents of this argument in fact produced their versions before Putnam, including Poincaré (1902/1952), Duhem, (1906/1954), Maxwell (1962) and Smart (1963). More recently, very sophisticated and precise formulations of this argument have been given by Boyd (1980; 1983; 1985; 1989), Brown (1982), Musgrave (1988), Lipton (1994) and Psillos (1999, esp. ch. 4), among others. Although there are, of course, differences in emphasis between the various formulations of this argument presented by these different authors, the common approach is one of appealing to realism as a means of *explaining* the success of scientific theories. Fine (1986a) accordingly dubs this the “explanationist defence of realism”.

It is worth making a few points of clarification before examining the NMA in detail. Firstly, notice that, while this argument takes the “success” of science as a basic premise, there are various possible definitions of “success”. All scientific realists are concerned with a theory’s *empirical* success, as opposed to, say, its success as an object of aesthetic admiration. Many contemporary authors interpret “success” as *predictive* success, but this interpretation is not the only one available. Boyd, for instance, understands overall empirical success, including explanatory success, to provide evidence for the truth of a theory. In Chapter 2 I will review arguments for and against according a special epistemic status to predictive success. Ultimately, I will endorse an account of theory confirmation which emphasises the importance of the unifying power of a theory, with novel empirical prediction serving merely as a reliable indicator that this underlying virtue is satisfied. For the purposes of the current chapter, however, our initial, admittedly vague, appeal to the notion of empirical success will suffice.

Secondly, it is worth noting that many anti-realists, including van Fraassen, accept that our best scientific theories are true insofar as they describe *observable* entities and processes. So the anti-realist can accept that the empirical success of our best theories is explained by their truth, although the claim under this formulation is relatively trivial. The realist, of course, wishes to claim that that our most successful scientific theories also give (approximately and/or partially) true descriptions of *unobservable* entities and processes. As Boyd puts it: “... non-observational terms in scientific theories should typically be interpreted as putative referring expressions" (Boyd, 1980, p. 613). In the remainder of this section, ascriptions of truth or approximate truth to a theory will therefore be taken to encompass the theory’s existence claims regarding, and descriptions of, both observable and unobservable entities.

Thirdly, there are in fact two distinct versions of the NMA on offer. Under one version, we take the overall “success of science” as our explanandum and argue that this lead to the conclusion that scientific methodology is capable, when carried out correctly, of arriving at approximate truth. Alternatively, we could argue from the empirical success of particular scientific theories to the truth of (at least some of) the claims contained within *these* theories. Reflecting upon this difference in emphasis Magnus and Callender (2004) distinguish between “wholesale” and “retail” arguments. I shall argue shortly that wholesale arguments are ultimately not tenable, though my reasoning differs somewhat from that of Magnus and Callender. It is also worth noting that the distinction between the two is not watertight. A retail version of the NMA could conceivably be applied successfully to all or most actually successful scientific theories, thus achieving a similar overall result to a successful wholesale argument.

Having addressed these clarificatory points, let us now examine the NMA in more detail. Although Putnam’s argument quoted above is relatively informal, we can formalise the argument using the template of “inference to the best explanation” (IBE). This form of argument has been discussed by Peirce (1958, book 2) under the heading of “abduction” and has been developed recently and in more detail by Harman (1965) and Lipton (2004). In general, IBE starts with some set of empirical facts as a premise. It then states that some postulate is the best explanation of these facts. Finally, it concludes that the postulate so identified is therefore very likely to be (approximately and/or partially) true. Note, however, that a postulate cannot be designated the “best explanation” simply on the grounds that it is most likely to be true; if we knew *that* in advance, we wouldn’t need to go through the tedious process of reasoning at all. Rather, “best” is defined in terms of “explanatory virtues”, such as simplicity, fit with the data, etc (Lipton, rather poetically, says that we choose the “loveliest explanation”).

This argument form is clearly not deductively valid. Indeed, if we interpret explanation as a type of deductive argument, then IBE is quite straightforwardly an example of the fallacy of affirming the consequent. Nevertheless, as with all ampliative arguments, we can formulate a deductively valid version of the NMA by adding additional premises. For clarity, I shall formulate both wholesale and retail versions of the argument:

<u>NMA (wholesale version)</u>

WP1: Science is empirically successful

WP2: We should believe the best ("loveliest") of the available explanations for the empirical success of science

WP3: There are exactly two available explanations for the empirical success of science: (a) that the methodology of science is reliably truth-conducive, and (b) that the success of science is the result of an unlikely coincidence (a "miracle")

P4: Positing the occurrence of an unlikely coincidence is never the best explanation if any other options are available

WC: *Therefore*, we should believe that the methodology of science is reliably conducive to ((at least approximate) truth.

<u>NMA (retail version)</u>

RP1: Theory T has been empirically successful

WP2: We should believe the best ("loveliest") of the available explanations for the empirical success of T

RP3: There are exactly two available explanations for the empirical success of T: (a) that T is (approximately and/or partially) true, and (b) that the success of T is the result of an unlikely coincidence (a "miracle")

P4: Positing the occurrence of an unlikely coincidence is never the best explanation if any other options are available

RC: *Therefore*, we should believe that T is (approximately and/or partially) true.

By formally stating both version of the NMA in this way, we can assess the overall soundness of these arguments by examining the soundness of their premises. In the following section, I shall therefore examine some objections to these premises. Along the way, I shall survey selection arguments, due to Fine and van Fraassen, which demonstrate that the wholesale version of the NMA is untenable (or, at least, even less tenable than the retail version).

## 3. Critical analysis of the NMA

Fine, (1984; 1986a; 1986b; 1991) has made several objections to the NMA. One objection questions the claim that science is "instrumentally successful", thus potentially discrediting both WP1 and RP1:

> "If, for example, we could examine the myriad attempts in laboratories around the world just (literally) yesterday to turn basic science to the production of a useful instrument, then, I think, we would find failure on a massive scale, and certainly not any overall success... For the application of science involves an enormous amount of plain old trial and error; hence, it always entails an enormous amount of error. I think a reasonable historical picture would be to draw each success as sitting on top of a great mountain of failure." (Fine, 1986a, pp. 152-153)

By "instrumental success", Fine seems to have it mind the ability to produce technological applications. This argument can, however, be very simply extended to include empirical success more generally. If we consider all the scientific hypotheses that are ever proposed, we will see that only a small proportion ever turn out to be empirically successful. It is only *after* a hypothesis is at least slightly successful that it will ever be formally reported, or even widely discussed.

So scientific methodology is not always, or even most of the time, effective at arriving at empirically successful theories. The generalised "success of science" that the NMA takes as a datum is in fact the result of a certain sort of confirmation or selection bias. We focus only the theories which do happen to be empirically successful, even though this result is not typical (see below for a detailed discussion of van Fraassen's selection argument against the NMA, which is closely related to Fine's). Note, however that, while this argument may be well-taken when directed against WP1 and therefore the wholesale version of the NMA, it is not obviously applicable to the retail version. RP1 in this retail argument refers to the success of a particular scientific theory, not to the enterprise as a whole. All parties to the debate, moreover, accept the existence of at least *some* empirically successful theories. I shall therefore take it that the soundness of RP1 is not contested by this argument.

Premises WP2 and RP2 essentially state that IBE is a truth-conducive (though not strictly truth-preserving) form of reasoning. Fine has also produced several arguments directed against this proposition. The first argument simply questions the claim that the success of science even *requires* an explanation. As Fine puts it: “[O]nce we have reached the age of reason we no longer accept just *any* explanatory challenge and history shows that, often enough, we are right not to do so” (Fine, 1991, p. 82). It is not altogether clear what history Fine has in mind here, but one plausible interpretation is that he is concerned about the proliferation of metaphysical claims on the basis of explanatory considerations. A line of argument can be reconstructed as follows. Empirical observation is the only truly sound rationale for coming to believe a given hypothesis. There is therefore never any reason to believe the truth of a hypothesis which is metaphysical in the sense that it is not empirically testable. Explanatory considerations can serve as a useful set of heuristics for devising new hypotheses to test, but they do not by themselves give us rational grounds for believing the truth of these hypotheses.

One possible response to this line of argument is to deny that scientific realism is a purely metaphysical thesis that is empty of empirical content. This thesis does make certain claims about what we should expect to see in the history of science, and these claims can be tested by a close examination of this history. This is a line of thinking I develop more fully in the remainder of this chapter, and so will not discuss it further in this section.

Alternatively, the realist may seek to provide some general positive argument for the truth-conduciveness of IBE. Fine, however, argues that any attempt to defend

IBE against an anti-realist opponent will inevitably be either question-begging or circular. Anti-realists, recall, do not accept the validity of even the “first-order” inference to the truth of a particular explanatory claim on the grounds that it is a good explanation of some set of empirical phenomena in science. So the anti-realist will be, if anything, doubly sceptical about applying a “second-order” inference to the truth of scientific theories on the grounds that this explains the success of these theories. As Fine puts it, “the realist is not free to assume the validity of any principle whose validity is itself under debate” (Fine, 1986a, pp. 160-161). If the realist does wish to construct an argument for the truth of scientific theories, she must “employ methods more stringent than those of ordinary scientific practice” (Fine, 1984, pp. 85-86). Anything less is simply question-begging.

Psillos, nevertheless attempts to provide a defence of IBE that is avowedly and unashamedly circular:

> “[The NMA] suggests that the best explanation of the instrumental reliability of scientific methodology is that background theories are relevantly approximately true. These background scientific theories have themselves been typically arrived at by abductive reasoning. Hence, it is reasonable to believe that abductive reasoning is reliable: it tends to generate approximately true theories.” (Psillos, 1999, p. 77)

So application of the second-order IBE, namely the NMA, tells us that first-order applications of IBE tend to give true beliefs. Hence, by induction over these instances of sound reasoning, we conclude that IBE is reliably truth-conducive. Psillos, taking his cue from various inductive defences of inductive reasoning (see Braithwaite 1953/1968, pp. 255-292; Black, 1958; van Cleve, 1984; Papineau, 1992; 1993, ch. 5), claims that the circularity of this argument is not a logical flaw. He distinguishes between “premise-circularity” and “rule-circularity”, and argues that only the former is “viciously circular”. A premise-circular argument is one in which the conclusion is identical to one of the premises of the argument. A rule-circular argument, in contrast, is one in which the reliability of some inferential rule is obtained as the conclusion of an argument which uses that very rule, but the reliability of the rule is not stated in the premises.

Psillos argues that rule-circularity is not viciously circular on the grounds that application of the rule does not require making any assumptions about whether or not it is reliable. He supports this claim by appealing to epistemic externalism, also known as reliabilism:

> “When an instance of a rule is offered as the link between a set of (true) premisses and a conclusion, what matters for the correctness of the conclusion is whether or not the rule *is* reliable... Any assumptions that need to be made *about* the reliability of the rule of inference, be they implicit or explicit, do not matter for the correctness of the conclusion. Hence, their defence is not necessary for the *correctness* of the conclusion.” (Psillos, 1999, p. 81, emphasis in original)

This type of externalist view has been advocated by Goldman (1979; 1986), among others. The opposed, so-called “internalist”, view is that any premise or form of inference used in an argument must be backed up by some sort of justification or reason before that argument can legitimately be accepted.

The fundamental difficulty with externalist approaches to knowledge is expressed by Papineau, as follows:

> “[R]eliabilism implies that whether or not we know will often hinge on matters, such as the reliability of some visual process, which lie quite outside our consciousness. But this seems to imply that we are at the mercy of nature, that we cannot do anything to affect whether or not we know. And this then makes reliabilist epistemology seem a quite different subject from the traditional version [of epistemology]. ... For surely a central concern of

> traditional epistemology was the normative question of what we should do in order to ensure that our beliefs are knowledge.” (Papineau 1993, ch. 5, s. 4)

We can all agree, in a metaphysical vein, that there is some fact of the matter about which of our epistemic practices are reliable. But lacking independent epistemic access to these facts, these metaphysical speculations do not help in the slightest in assessing whether a given practice *is* in fact reliable. Because it is not in any way action-guiding, therefore, the approach proposed by Psillos is therefore unsatisfactory.

It is, in any case, deeply implausible that the proponent of IBE will *ever* be able to devise any argument for it, internalist or externalist, that will persuade a sceptic. Just as, following Hume, we should regard it is deeply implausible that the proponent of inductive reasoning can convince a sceptic. And indeed, as argued by Carroll (1895), even *deductive* reasoning is not immune to challenge by a sufficiently determined (though probably only hypothetical) sceptic[1]. At best, the advocate of IBE can show that this form of argumentation gives rise to an overall coherent system of propositions; this, of course, being simply a different way of describing the circularity of Psillos’ argument sketched out above. The best case scenario for the proponent of IBE, is that we conclude that there are no decisive rational arguments *against* accepting it. This allows for a sort of ‘voluntarist’ attitude towards this mode of reasoning. That is, to either accept it or reject it is to enter into one particular ‘style’ or ‘practice’ of rationality. Under the voluntarist picture, there are no rational arguments that the proponent of either style could offer that could hope to convince her opponent. As indicated above, however, this is not necessarily a terrible position for the IBE advocate, as inductive and deductive reasoning would seem to be in the same boat.

[1] This sceptic (Carroll’s tortoise) may ask why, given the premises A and A→B, he ought to believe B. It seems the only possible response is to add another premise “If A and A→B, then B”. But then the sceptic will still doubt that the conclusion follows from these *three* premises, and so a regress ensues.

van Fraassen (1989, pp. 142-150), however, has an argument which threatens to confound IBE even on its own terms. Suppose we accept that it is possible to rank a given set of theories according to some criterion that correlates with truth, so we can always select the theory that is closest to the truth. But this ranking always takes as its input only those theories that have actually been proposed. And these may all be quite far from the truth. So IBE may only give us the “best of a bad lot”. While it will get us progressively closer to the truth, it never justifies the assertion that a given hypothesis *is* (even approximately and/or partially) true.

There are various possible replies to the “bad lot” argument, several of which are outlined by Lipton (1993; 2004, pp. 151-163) and Psillos (1999, pp. 208-214). The most promising of these is Lipton’s proposal to “eliminate the gap between comparative and absolute ranking... by exhaustion” (Lipton, 1993, p. 94). That is, if the set of theories from which we draw contains every logically possible option, and our ranking reliably picks out the theory closest to the truth, then we will be able to draw a true theory from this set. This argument is cogent, so far as it goes. The obvious problem with it is that no actual case of theory choice in science takes the entire space of logically possible theories as input. In real cases, there are *always* logically possible theories that have not been considered.

Lipton puts a great deal of weight on the one exception to this general problem, the case where a set of proposed hypotheses consists of only a single hypothesis and its negation. For example, the claim that some entity X is the cause of a particular phenomenon and the claim that it is not together comprise such a set. If, *ex hypothesi,* we have the means to pick from a set the hypothesis that is closest to the truth, then we will always pick the true hypothesis from such a set. Again, however, it is extremely unclear how this argument could be carried over to real cases, as very few genuine scientific hypotheses consist of single isolated propositions. Moreover, as Ladyman (in Ladyman & Lipton, 2006) points out, the goal of the NMA is merely to select theories that are (at least) *approximately* true. Even given a case of mutually contradictory theories, Lipton’s reasoning does not help with this more modest goal. The approximate truth of A does not entail the falsity of ¬A. They can both be approximately true together, even if they are strictly logically incompatible.

The difficulty of constructing a general defence against the bad lot objection is perhaps not so surprising when one recognises that this objection quite closely resembles the both the old and intractable problem of underdetermination of theory by evidence and Stanford's (2006) more recent "problem of unconceived alternatives". All of these problems arise because there is always a logical possibility (or, as some would argue, near-certainty) that, for any given theory which is accepted by scientists, there other theories which have not been considered and yet which would be regarded as better confirmed by all the available evidence if they were.

Proponents of the underdetermination problem are entirely correct in arguing that the logical possibility of another theory that is both radically dissimilar to and better-confirmed than the one in question can never be eliminated. However, in this context, the NMA can be viewed as an attempt to show that this possibility is remote. Interestingly, it attempts to do so by offering a putatively exhaustive set of explanations for a putatively accepted fact, namely that some theory (or science in general) is successful. Focussing on the retail argument (though the same point applies, *mutatis mutandis*, to the wholesale argument), RP3 asserts that there are only two possible explanations for the success of a theory, namely (a) that the theory is (approximately and/or partially) true; and (b) that it is radically false (i.e. not even approximately and/or partially true) but successful because of a lucky coincidence. Explanation (b) is intended as a catch-all that is true just in case any of the alternatives to the favoured theory is in fact true. P4 then simply denies that (b) can be true, leaving the truth of the successful theory as the only explanation. If sound, this argument defeats both the problem of underdetermination and the bad lot argument.

This brings us quite naturally, then, to a closer examination of WP3 and RP3. These premises each assert that there are only two possible explanations for the success of science, or the success of a particular theory, respectively. These premises can therefore be undermined simply by introducing plausible alternatives to these two possibilities. van Fraassen (1980) offers a specific alternative, as follows:

> "In just the same way, I claim that the success of current scientific theories is no miracle. It is not even surprising to the scientific (Darwinist) mind. For any scientific theory is born into a life of fierce competition, a jungle red in tooth and claw. Only the successful theories survive—the ones which in fact latched on to actual regularities in nature." (*op. cit.*, p. 40)

Accepted theories tend to be empirically successful, in other words, because scientists are only willing to accept the small number of theories that happen to be successful. This "selectionist argument" and some of the responses to it are ably summarised by Wray (2007, 2010).

Lipton ( 2004, pp. 193-195) has shown van Fraassen's selectionist argument to be unsatisfactory. Firstly, he points out that the proposed explanation is not *incompatible* with the explanation that the theory in question is true. It may be that the theory is empirically successful, and so is selected, precisely because it is true. But the anti-realist could concede this and still argue that the selection mechanism is sufficient to explain the phenomenon in question, so any additional explanation is superfluous. The selectionist explanation, in Lipton's words, "pre-empts" the explanation in terms of truth.

Lipton, however, gives two further arguments against the selection explanation. Referring to Nozick (1974), p. 22), he notes that a selection mechanism can explain why all members of some collection have a certain property, but will not explain why each of them has that property. For instance, we can explain why all members of a club have red hair by reference to the fact that the club only admits those with red hair. But referring to this policy does not explain why each of these individuals has red hair; an explanation of this fact would presumably require knowledge of the genetic makeup of each of the individuals concerned, or something along these lines. Analogously, we can concede that all the theories that we accept are empirically successful precisely because we accept only successful theories, and still demand an explanation for why each theory is successful. Notice, incidentally, that this argument amounts to a concession that the selectionist argument works against a *wholesale* version of the NMA. Indeed, in this respect, van Fraassen's argument is substantially identical to Fine's argument given in the first paragraph of this section.

So the question is whether the selectionist argument can undermine a *retail* version of the NMA. This brings us to Lipton's second criticism of the selectionist argument, that it is ambiguous on the notion of "empirical success". The selectionist argument seems perfectly able to explain why we now have theories that successfully describe empirical phenomena that we have observed *in the past*. But it is unable to explain why some theories are able to successfully *predict* phenomena that were unobserved at the time the theory was initially accepted. Van Fraassen says that empirically successful theories have "latched on to actual regularities in nature", but given the overall project of constructive empiricism, it is clear he means regularities in *observable* phenomena. Setting aside the problem of induction, we can assume that all parties to this debate accept that statements of empirical regularities can legitimately be projected into the future to give predictions. But this type of low-level induction only works for phenomena in the same broad class as those observed before. The prediction of any new *type* of phenomenon inevitably appeals to claims that the theory in question makes about *unobservable* entities or processes. An explanation by reference to truth explains why theories selected for their success in describing one set of empirical phenomena can successfully predict a different set of phenomena, whereas the selectionist explanation cannot. Thus the selectionist explanation is not a plausible alternative for the explanations suggested in RP3.

This, finally, brings us to a consideration of premise P4, which denies the explanatory legitimacy of an appeal to lucky coincidence. There are two main objections to this premise. van Fraassen argues simply that coincidences are frequently quite explicable:

> “[I]t is illegitimate to equate being a lucky accident, or a coincidence, with having no explanation. It was by coincidence that I met my friend in the market—but I can explain why I was there, and he can explain why he came, so together we can explain how this meeting happened. We call it a coincidence, not because the occurrence was inexplicable, but because we did not severally go to the market in order to meet. There cannot be a requirement upon science to provide a theoretical elimination of coincidences, or accidental correlations in general, for that does not even make sense.” (van Fraassen, 1980, p. 25)

This, surely, is correct as far as it goes. But Forster (1988a) and Ladyman (2002, p. 216) argue that it is not a legitimate analogy to the empirical success of a scientific theory. They claim that *repeated* coincidences of this nature eventually rule out coincidence as an explanation. As Forster puts it: “[I]f each man went to the market at irregular times just once a week and kept on meeting week after week, then we would suspect that there is something more to be said” (*op. cit.,* p. 544).

While Forster and Ladyman are correct in stating that van Fraassen’s analogy is spurious, there is a stronger reply available. To introduce this reply, let us briefly consider the lottery paradox (Kyburg, 1961, p. 197). Suppose exactly one ticket for the lottery will be selected, randomly, as the winner. But a large number of tickets have been sold, so it is extremely unlikely that any given ticket will be the winning ticket. And yet *some* ticket has to win. So an extremely unlikely event is certain to occur! One way of resolving this paradox is to notice that there is an ambiguity in how we have specified the event – it is unlikely for a *specified* particular ticket to be chosen, but it is not unlikely (indeed, it is certain) that *some* ticket will be chosen. The point is that it does not require any great predictive capacities to say that *some* ticket will win, or that a given person will *at some point* run into *some* friend in the market, or even that *some* theory will repeatedly predict the results of measurements. Given the structure of the respective event spaces, these events are all in fact quite likely. A special explanation is only required if a particular ticket (specified in advance by a guru, say) wins the lottery, or the theory accepted by scientists on separate grounds predicts the results of an experiment. In fact Forster and Ladyman's intuition that repeated occurrences of a given type of event require a special explanation probably stems from the fact that the first occurrence generates a tacit 'prediction' that there will be a recurrence. And it is the satisfaction of this prediction that demands an explanation.

The final argument to address is the "base-rate" argument given by Howson (2000, pp. 52-54) and also discussed recently by Lipton (2004, pp. 197-198) and Magnus and Callender (2004). Before beginning in earnest, it is worth noting that Howson (a good Bayesian if ever there was one) thinks of the NMA in probabilistic terms. So a successful version of the NMA would be able to assign a high probability to the hypothesis that a theory is (approximately and/or partially) true. The base-rate argument purports to show, contra P4, that even assuming a theory is empirically successful, it is not legitimate to infer that the probability of this theory being successful by "lucky coincidence" is lower than that of it being true. Suppose that all true theories are successful. However, suppose also that some number of false theories are successful. The information that a theory is empirically successful therefore tells us only that it is in the class containing true theories and successful false theories. Because there are many more false theories than true theories, if even a tiny proportion of false theories are successful, there may nevertheless be sufficiently many of them to vastly outnumber the true theories in this class. Therefore, although discovering that a theory is successful certainly should increase the subjective likelihood of it being true, it doesn't necessarily warrant the belief that it is *likely* to be true (even approximately and/or partially). To make such an inference ignores the "base rate", or the proportion of 'good' cases in the class from which samples are being drawn.

The base-rate objection, incidentally, should not be understood as claiming that empirically successful theories are likely to be *false*. Its purpose is simply to draw attention to the fact that we have *no idea* what the correct base-rate, or prior probability, is of a given theory being true. To put this another way, we do not know anything about the structure of the 'population' from which a particular theory is being selected. One way for the realist to respond to the base-rate objection, then, would be to provide some argument for the claim that the theories typically considered by scientists have a relatively high probability of being true. Note that this probability still need not even approach unity – it only needs to be substantially higher than the probability of a given successful theory being false. If it is legitimate to assume a relatively high prior, then the additional evidence that a particular theory is empirically successful provides good evidence that this theory is likely to be true.

One example of such an argument is given by Psillos (2006), who argues that it is legitimate to set priors by applying the principle of indifference to an outcome space consisting of only two propositions: the theory in question (T) and its negation (¬T). The result of this calculation is that the prior probability of both T and ¬T is set to 0.5. With such a distribution of priors, the observation of evidence compatible with T can legitimately lead to the judgement that T is likely to be (approximately and/or partially) true. The obvious problem with this argument, one that at least potentially affects any application of the principle of indifference, is that the outcome space is blatantly gerrymandered to reflect the desired outcome. It is hard to see how the truth of one theory and the truth of all the other possible theories can legitimately count as one outcome each. Howson's approach is, in this respect, far more defensible. Notice that the base-rate argument also tacitly applies the principle of indifference, except that it imposes a *prima facie* much more neutral classification of outcomes, treating the truth of every possible theory as an outcome, and accordingly arguing that the initial probability should be distributed uniformly across them. Howson's conclusion that we cannot state any particular value for the prior probability of T (even a very low one) follows relatively straightforward from this procedure plus our ignorance about the size of the population, and what proportion of the theories it contains is true.

Worrall’s (2005) resolution (note, not “solution”) to this problem is slightly more satisfactory. He connects the NMA, and IBE more generally, explicitly to the problem of induction. However much we dress each up in deductive or probabilistic language, he argues, each form of reasoning remains fundamentally the application of a pre-theoretical intuition. And, while it is certainly possible that these intuitive judgements turn out to be systematically mistaken, the “reasonable default position” is that inductive generalisations and judgements about the (approximate) truth of theories turn out to be correct. This “default position” on NMA can be interpreted as a tacit assumption that the theories found in mature sciences enjoy high priors, just as the intuitive application of inductive reasoning can be interpreted as tacitly presupposing the ‘uniformity of nature’ (or something similar). And both forms of reasoning can quite legitimately be rejected by rejecting these assumptions. But they can legitimately be accepted also.

This, then, returns us to the sort of ‘voluntarist’ attitude towards IBE suggested above. There is no rational argument that will convince the committed sceptic of IBE; but it is also not *irrational* to apply these intuitive forms of reasoning. Rather, the adherent and the sceptic inhabit different modes or ‘styles’ of rationality. In the remainder of this thesis, therefore, I will not take a position on the soundness of the NMA. I will, in later chapters, make some arguments about what other positions realists ought to accept *given* that they accept the NMA. But I will not try to judge whether, as a metaphysical position, any given form of realism is rationally justified. It is sufficient for the arguments that follow that scientific realists are committed to the NMA under some formulation.

## **4. Naive realism and the pessimistic meta-induction**

A large part of the argumentation in this thesis is directed instead towards a particular consequence of the NMA, specifically a commitment to substantial continuity in the history of science. Consider a scientific theory that is uncontroversially empirically successful at some period in history, past or present. If the NMA is valid, then this theory is (approximately and/or partially) true. However, generations of scientists following this period will almost certainly arrive at some successor theory which is also successful. Applying the NMA again, it follows that this successor theory is also (approximately and/or partially) true. To say that a theory is (approximately) true is to say that it (approximately) accurately describes those parts of the mind-independent world it purports to describe. Any two theories or models which approximately accurately describe the same part of the world must be substantially 'similar' to each other in important respects. There should therefore be substantial continuity between a successful theory and any later theories in the same domain that are also successful (The notions of "similarity" and "continuity" are left deliberately vague here. They will be defined with much more precision in later sections of this thesis). Adding this corollary explicitly to the commitments of scientific realism, "naive realism" (NR) is defined as follows:

> If a theory is empirically successful, then we have good reason to believe (i) that it is (approximately and/or partially) true and (ii) that there will substantial continuity between it and any successor theory.

As any regular reader of philosophy knows, any position labelled "naive" is surely doomed. And so it is here. In the hands of anti-realists, the commitment to continuity is wielded as an argument against scientific realism. This is the so-called "pessimistic meta-induction" (PMI). An early version of this argument is found in Poincaré:

> "The ephemeral nature of scientific theories takes by surprise the man of the world. Their brief period of prosperity ended, he sees them abandoned one after the other; he sees ruins piled upon ruins; he predicts that the theories in fashion today will in a short time succumb in their turn, and he concludes that they are absolutely in vain." (Poincaré, 1905/1952, p. 160)

It has also been suggested more recently by Putnam:

> “What if all the theoretical entities postulated by one generation (molecules, genes, etc., as well as electrons) invariably 'don't exist' from the standpoint of later science? This is, of course, a form of the old sceptical 'argument from error' - how do you know you aren't in error now? But it is the form in which the argument from error is a serious worry for many people today, and not just a 'philosophical doubt'. One reason this is a serious worry is that eventually the following meta-induction becomes overwhelmingly compelling: just as no term used in the science of more than fifty (or whatever) years ago referred, so it will turn out that no term used now (except maybe observation terms, if there are such) refers.” (Putnam, 1978, pp. 24-25)

However, the best-known version of the PMI is Laudan’s, perhaps because he offers a concrete list of theories which he claims are “both successful and (so far as we can judge) non-referential with respect to many of their central explanatory concepts” (Laudan, 1981, p. 33). In each case, the argument begins with the claim that most past scientific theories that were highly empirically successful are nevertheless, by the lights of contemporary science, not even approximately true. It then inductively generalises from the falsity of these empirically successful theories in the past to the falsity of our *current* empirically successful theories. Hence, realism is refuted.

The PMI can be written down more formally, as follows:

P1: For most cases in the history of science where, at different times, different theories have purported to describe the same set of empirical phenomena, there is radical *discontinuity* between earlier and later theories. This is the case even where both theories concerned are empirically successful.

C1: Thus, in these cases, at least one of the older or the later theory must *not* be even approximately true.

P2: (Assume that) theories that are currently empirically successful are approximately true.

C2: So, any past theory which bears the relation of radical discontinuity to our present theories cannot be even approximately true.

C3: Therefore most empirically successful theories in the past have been not even approximately true.

This intermediate conclusion establishes an inductive base of historical theories that we should regard as false, despite their success. To perform an inductive inference requires as an additional premise:

P3: Present scientists are in relevantly similar epistemic circumstances to their scientific ancestors.

Eventually producing the final conclusion:

C4: Therefore, by inductive generalisation from past theories, theories that are currently empirically are very probably not even approximately true.

Thus, argues the anti-realist, scientific realism is refuted. Notice that, formulated above, this argument assumes at P2 the crucial realist claim that is later refuted in the conclusion. It therefore has the form of a *reductio.* Various criticisms of this argument can be assessed by careful examination the soundness of the various premises, and the validity of the inductive step. To start, although it is a trivial point, it is worth mentioning that no realist critic of the PMI attacks P2. This claim, after all, is precisely what the realist intends to defend.

Premise P1 has been attacked on several grounds. The most prominent response will be elaborated in detail in the following section, under the heading of “partial realism”. Partial realism claims that, despite all superficial appearances, there is in fact generally substantial continuity between theories which are empirically successful in the same broad domain at different times. This is because the parts of theories actually “essential” for their empirical success tends to be conserved in successor theories. Note, however, that this response effectively concedes the point that naive realism – the view that empirically successful theories are approximately true in every respect – is susceptible to the PMI.

Notice that P1 claims that *most* of the relevant cases of theory change are characterised by radical discontinuity. For simplicity, assume the validity of Reichenbach’s (1949, p. 446) ‘straight rule’ of inductive generalisation. That is, that the proportion of cases of interest in the overall population (in this case, empirically successful theories that turn out to be false) will roughly approximate the relative frequency of cases of interest in the sample observed so far. So, the claim that *most* cases of theory change are characterised by radical discontinuity is necessary to sustain an inductive generalisation to the effect that our currently successful theories are “very probably” not approximately true. The realist, in response, might however grant that *some* proportion of successful theories fail to be even approximately true (say those on Laudan’s list). But she may deny that most or even many do. Perhaps, she might argue, the cases cited by Laudan and others are actually quite unusual, and the vast majority of empirically successful theories turn out be approximately true. If this is the case, one might inductively infer that our current empirically successful theories are very probably approximately true. So the PMI can be turned out its head to create an *optimistic* induction. This idea is discussed in more detail in section 7.

P3 has recently come under attack by Roush (2009). She concedes that our scientific predecessors were epistemically unreliable (i.e. often believed in false theories). But Roush points out, correctly, that an inductive inference can be undermined by pointing to some relevant factor that differs between the inductive base and the case for which a conclusion is to be drawn. And she argues that the fact that current methods of scientific investigation differ from (and are superior to) those in the past is a sufficient reason to regard the inference as unsound. The obvious rejoinder to this line of thinking, however, is raised by Roush herself:

> “The pessimist may reasonably protest that even if we show our methods to be different from our predecessors’, many of our predecessors also had methods that were different from their predecessors’. A lot of good it did them since they ended up wrong a lot of the time.” (*op. cit.*, p. 53)

She responds to this concern by stating that a difference in method is *always* sufficient to undermine any inductive generalisation about reliability from one situation to the next. Indeed, she goes so far as to say that “our predecessors’ predecessors’ failures did not render our predecessors’ confidence unjustified either” (*op. cit.*, pp. 53-54).

It is extremely unclear, however, how Roush can hope to justify this assertion. Change can itself surely be intelligibly understood as part of the inductive base for an inference, so long as it is systematic or regular in some way. For instance, human beings have attempted to devise perpetual-motion machines for centuries. And they have always attempted to incorporate the most advanced available technology in doing so – water wheels, clockwork, magnetism, electricity, etc. And yet they have never succeeded. Granted, there is now a well-confirmed physical law – the second law of thermodynamics – which states that such a device is impossible. But a large part of why we regard this law as “well-confirmed” is precisely because it is the result of an inductive generalisation from this abundant experience of failure! Analogously, suppose it is the case that, over the whole history of science, scientists have continually applied various methods to devise scientific theories, and these theories have consistently turned out to be false by the lights of later theories. So, although the details change, current scientists are in a very real sense simply doing what scientists have always done, and the pessimistic induction still bites.

Another objection to the PMI has been suggested by Lewis (2001). This objection has gained some prominence recently, with Magnus and Callender (2004) declaring that the sorry state of the PMI (and the NMA) was sufficient to provoke "*ennui*" about the entire scientific realism debate. Lewis' argument makes use of the "base rate" reasoning discussed in section 2 above. He defines a "reliable test" as one where the rate of both "false positives" and "false negatives" is low. In the context of arguments for realism, such a test has only a small probability of judging a theory as true if it is in fact false, and vice versa. The core of his argument is then as follows:

> "[Laudan's] claim is that ... theories which are successful but false ... outnumber ... theories which are successful and true ... But as we saw above, it would be a fallacy to conclude from this evidence alone that success is unreliable as a test for truth. If true theories are relatively rare, then it is only to be expected that false positives outnumber true positives, even if success is a reliable test for truth." (Lewis, 2001, p. 376).

So Lewis concedes that the majority of theories we infer to be true on the basis of their may well be false. The point, for him, is that this is perfectly compatible with this form of inference been extremely reliable or truth-tropic.

But notice that this is substantially the scenario outlined under Howson's base-rate argument *against* realism in the previous section! And for good reason – the realist's primary goal is to demonstrate that that our best scientific theories are probably (approximately and/or partially) true. Suppose you are told (let us assume correctly) that you have a genuine talent for beating the odds when it comes to picking lottery tickets – instead of ten million-to-one, you typically face only million-to-one odds. But this is hardly grounds for quitting your job on the basis of the ticket you currently hold. Similarly, it is surely cold comfort for the realist to be told that, although our most empirically successful theories are probably false, the probability of them being true is better than that for scientific theories in general. So Lewis' argument doesn't defeat the PMI. It does not, for that matter, make any point that Laudan would find surprising or troubling, since the latter was concerned to argue against the claim that our best scientific theories *actually are* true.

## 5. Partial realism and deflationary realism

It seems, then, that both the NMA and the PMI have at least some force. Many proponents of realism have responded to this situation by seeking, in Worrall's (1989a) words, to have the "best of both worlds". They respect the force of the PMI by rejecting the "naive realist" position outlined above. But they respect the force of the NMA by seeking to be realists about *parts* of scientific theories. They do so by distinguishing between the elements of a theory "*essential"* or "*inessential*" for its empirical success, then arguing that only the latter are discarded in instances of theory change. There are several names for this strategy, including "preservative realism" (Chang, 2003), "selective confirmation" (Stanford, 2003a; 2003b; 2006), "localized realism" (Elsamahi, 2004) "selective scepticism" (Chakravartty, 2007; although he at times uses this term sufficiently broadly that it also includes constructive empiricism) and "selective realism" (Cei, 2009; Saatsi in Saatsi, *et al.*, 2009).

My own characterisation of this strategy places more emphasis on theoretical continuity than these authors do. I therefore coin the new term "partial realism" (PR) to describe the following view:

> If a theory is empirically successful and some element of this theory is essential for that success, then we have good reason to believe (i) that *this element* accurately describes a corresponding feature of the world and (ii) that this element will be preserved in successor theories.

PR is a relatively abstract position, and is intended to capture various more specific views that have been advocated in recent years. These views differ largely in how they characterise the key term "essential" in PR. The structural realist claims it is theoretical "structure" that is essential to prediction and is preserved in later theories; Kitcher assigns this status to “working posits”; etc (see Chapter 3 for a detailed discussion of these alternatives). There are also serious debates to be had concerning the meaning of “empirical success” and “preserved”. The term “preserved” will be discussed later in this chapter, but the other vague terms require more detailed exposition, and so will be discussed in subsequent chapters.

Whereas both naive and partial realists argue from the success of theories to their truth and thence to the continuity of science, it is also possible to construct a plausible argument in support of continuity as a thesis in its own right. A particular version of this thesis advocated here, and I have given it the name ‘deflationary realism’ (DR). It is defined as follows:

> If a theory is empirically successful and some element of this theory is essential for that success, then we have good reason to believe that this element to be preserved in successor theories.

DR departs from PR only in that it omits the traditional realist commitment to truth. It should be emphasised again that this thesis is officially neutral on the soundness of the NMA and the (approximate and/or partial) truth of successful theories. In fact, there are several positive reasons for avoiding a commitment to these issues one way or the other. The most important is that, like many metaphysical issues, it is unclear what hinges upon a theory being true, as opposed to empirically adequate, etc. This issue is crystallised by the question “Is there some difference in how a realist and an anti-realist would do science?” There are both realists and anti-realists who think the answer to this question is a decisive “no”. Recall, for instance, the criticism of realism attributed to Fine in section 3, that it is utterly without empirical content and so ought not to be the subject of argument.

Many realists and anti-realists will agree with Fine’s descriptive assessment, but disagree that it represents a criticism of realism. The dispute at hand is, for them, metaphysical but nevertheless interesting. No position will be stated here on whether purely metaphysical questions are indeed of philosophical interest. It will be argued, however, that scientific realism is simply *not* a purely metaphysical position. The realist does have at least one practical commitment that the anti-realist does not (or, at least, need not) have. Specifically, the realist is committed to the continuity of science, as demonstrated in section 2. It is the nature and soundness of this commitment that will animate discussion in the remainder of this thesis.

It should be clear exactly what is “deflationary” about deflationary realism. In general, a deflationary view seeks to preserve the practical or functional dimensions of some concept, while avoiding the metaphysical commitments that often accompany philosophical views of it. For instance, while the traditional philosophical account of truth cashes out the concept as “correspondence with reality” or something similar, the deflationary view is much more interested in what ordinary speakers hope to accomplish when they use the term “truth”. Similarly, DR declines to draw any inferences about the nature of the mind-independent world, but is interested in how a ‘realist attitude’ to some part of a scientific theory may shape scientific practice. It should be noted, however, that although the focus throughout this thesis will be on the sort of scientific conservatism embodied in DR, it is not necessarily the case that this is the *only* such consequence of realism. A deflationary view is also not *incompatible* with more metaphysically ambitious positions. The more metaphysically inclined reader is therefore welcome to take the following arguments surrounding DR on their own terms, and supplement them with additional metaphysical commitments if she so desires.

The realist may well wish to deny the deflationary realist the right to be called a “realist”, with or without the qualifier. She might argue that a commitment to the truth of theories is *just what it is* for any position to count as realist, and dropping this commitment results in anti-realism, whatever else is advocated or denied. This is a dispute over the proper use of a concept, rather than an argument over a philosophical thesis, and consequently intuitions over the proper resolution will diverge wildly. A similar dispute arises in relation to the deflationary account of truth, which many argue is not an account of truth at all. I favour the term "deflationary realism" because, as argued briefly above, DR captures an extremely important commitment of the realist position. Yet no substantive point in this essay rests on whether DR is itself understood as a realist position. Thus, if the reader would prefer, he or she is welcome to read "commitment to the continuity of science" in place of "deflationary realism" hereafter.

As a way of outlining the arguments given in the remainder of this chapter, I will briefly sketch out some of the logical relationships between these views. Firstly, it is worth noticing that that NR logically entails PR, which in turn logically entails DR. The NMA provides an argument in favour of both NR and PR from the empirical success of particular theories. However, the NMA is blocked in respect of NR by the PMI, leaving it to support only PR. Separately, we might make an 'optimistic induction' from theoretical continuity (note, *not* empirical success) to DR (see section 7). Finally, we might construct an "argument from continuity to truth" (ACT, see section 9) from DR to PR.

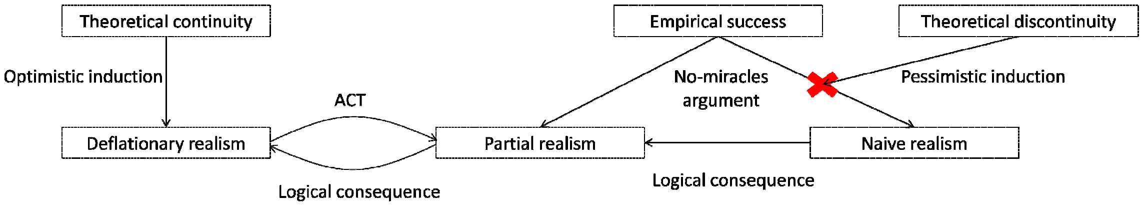

**Figure 1.** The logical relationships between the various forms of realism discussed in this chapter.

Some arguments have been given to the effect that PR is a more defensible position than NR. No arguments will be given, however, that aim to demonstrate the superiority of DR over PR (or vice versa). Rather, I will suggest that we merely *entertain* DR, as a means of providing insight into the debate between scientific realism and anti-realism. The investigation of DR offers several advantages over the investigation of more metaphysically inflationary forms of realism. Firstly, since both NR and PR are committed to DR, simple logic states that a refutation of the latter also counts as a refutation of both of the former, although the converse is not true. This is a major point of appeal for DR: it gives us a wedge into the realism/anti-realism debate. And, since it is not committed either way on the most contentious metaphysical questions, the business end of this wedge is relatively sharp. Secondly, because DR makes no metaphysical claims, it is in principle as acceptable to the anti-realist empiricist as it is to the realist. Thirdly, DR expresses one of the key practical commitments of scientific realism. Therefore, even if they differ on metaphysical questions, the empiricist who is willing to accept DR has much more in common with the realist than either of them may realise.

## **6. What elements are preserved?**

One may well wonder how a commitment to the continuity of science can be described as a “practical commitment”. The answer to this is that the thesis of the continuity of science underpins an important methodological standard, namely use of the “general correspondence principle”. This states that “... any acceptable new theory L should account for the success of its predecessor S by ‘degenerating’ into that theory under those conditions under which S has been well confirmed by tests” (Post, 1971, p. 228).

Applied consistently, the general correspondence principle serves as a significant constraint on scientific theorising. Indeed, Post recommends that we “consign to the wastepaper basket” any prospective successor theory L which fails to explain the success of S. The rationale for this, presumably (Post never states it explicitly), is that pursuing this new theory any further is virtually guaranteed to be a waste of a scientist’s time and resources. Put more positively, a commitment to the principle can help to limit the seemingly limitless field of options.

Post’s view is, in general, extremely friendly to the partial or deflationary realist, as he also attempts to distinguish essential from inessential parts of a theory. He suggests that we can distinguish the parts of a theory that are well-confirmed (designated S*) by carefully examining “vertical” and “horizontal” divisions within it. The vertical divisions distinguish those parts of a theory which are successful in their intended empirical domain from those that are not. His own example is the phlogiston theory of chemistry, which he argues effectively accounted for patterns in chemical reactivity, and yet was totally inadequate as an explanation of colour differences between substances.

The “horizontal” divisions within a theory correspond to progressively more general or abstract statements comprising the “higher” levels, with “[t]he well-confirmed part of a theory extend[ing] only up to a certain level” (*op. cit.*, p. 229). The lower-level parts which are well-confirmed provide most of the empirical content of the theory, whereas the higher-level parts “interpret” the empirical regularities. The “patterns”, particularly classifications, at the lower levels tend to survive scientific revolutions, whereas interpretations at the higher levels tend not to. Post further elaborates this idea with the aid of an analogy. He, like Kuhn (1962/1996, esp. ch. 9), compares a scientific revolution to a political revolution. However, unlike Kuhn, he emphasises that an unstable series of governments might be underpinned by a stable civil service which carries on without significant disruption. Analogously, Post claims that much of the workaday machinery of a scientific theory will be preserved over the course of a revolution.

This general picture is certainly suggestive, and Post does back it up with some examples. However, these examples appeal simply to the reader’s intuitive judgement that elements have indeed been preserved across theory change. Post does not, in other words, give a precise general account of which elements of a theory are to be counted essential. Various more precise accounts of “essentialness” will be examined in Chapter 3, and I will offer my own views. However, the remainder of this section consists simply of a brief taxonomy of the different cases of theoretical elements being preserved, using mainly examples given by Post himself. At least five different kinds of cases are distinguishable, and these will be listed in order of (roughly) increasing abstraction of the theoretical elements preserved.

The least abstract sort of case is where predecessor theory S and successor theory L correspond in the predicted values for a particular set of measurable quantities. One example of this may be the predicted apparent positions of the planets in the Ptolemaic versus Copernican astronomical systems. Each system reproduces roughly the same data set, but even the basic models underpinning them differ quite significantly. As Post: “Astronomy ... has retained a core of kinematics observed in the heavens while attributing this to gods, holes in crystal spheres, or gravitating masses” Post, 1971), p. 238).

The second sort of case involves the preservation of a low-level ‘empirical’ model from one theory to the next. This is more abstract than the previous kind of case, since what is preserved is not merely a particular set of results, but a general model with free parameters that may be fixed so as to provide predictions for a variety of particular cases. Post’s example in this category is the Balmer formula for predicting the frequencies of atomic spectra, which is found in both Bohr’s original quantum theory and in the later quantum electrodynamics.

The third sort of case involves “mid-level” theoretical posits. One example, not mentioned by Post, is the general transverse wave equations for light. These are preserved from Fresnel’s theory to Maxwell’s theory, although they are interpreted first as describing perturbations in a mechanical ether and later as describing oscillations in a free-standing electromagnetic field (see Worrall, 1989a). It is worth noting, moreover, that some of the lower level consequences of these equations are also preserved, including Fresnel’s equations for refraction and reflection at the interface between two media.

The fourth sort of case is what Post refers to as “inconsistent correspondence” (*op. cit.*, p. 232). This sort of case is very similar to the second and third sorts described above, except that here the equation or pattern posited by S* is *strictly* inaccurate by lights of L, but *approximately* accurate under some intuitively reasonable limiting conditions. The venerable example, cited by Post, is that the equations of Newtonian classical mechanics are recoverable from Einstein’s relativistic mechanics as the ratio *v/c* approaches zero, where *v* is the speed of the object we are considering and *c* is the speed of light in a vacuum. Of course, the speeds of actual objects are certainly not always near zero. And yet the circumstances in which classical mechanics have been observationally confirmed all involve objects moving at speeds which are very small relative to that of light, so the predictions of classical mechanics are *quantitatively* similar to those of relativistic mechanics.

The fifth kind of case involves the preservation of some pattern of classification. Post gives two examples of this sort of case. The first, as already discussed, is the classification of chemical species by their degree of phlogistication (this particular case study will be examined in much greater detail in Chapter 5). The second involves the periodic system, whereby elements are arranged into groups with similar chemical reactivity, these properties recurring after regular intervals. When originally devised, the periodic law ordered elements by weight. However, in later versions, the elements are ordered by atomic number (a concept not available to the original framers of the system). The correlation between these two quantities is sufficiently close that the predicted patterns in chemical reactivity are, with a few interesting exceptions, preserved between system of ordering and the other.

What is common to all these types of cases? As Post emphasises, in each case the successor theory L is intuitively able to “account for”, that is to say *explain*, why the earlier theory S was successful. The relationship between a later theory and an earlier one is however, according to the partial realist, stronger than that of mere explanation. This is for the simple reason that a theory which is largely true could quite satisfactorily explain the success of one which is entirely false except in respect of its observational consequences. But the partial realist claims that the essential elements of *all* empirically successful theories (at least approximately) accurately represent corresponding entities or processes in the mind-independent world. So two theories which are successful in respect of the same empirical phenomena must accurately represent the same entities, processes and/or properties. There are therefore two potential candidates for the type of relationship that, according to the partial realist, holds when the essential elements of a given theory are “preserved” in its successor theories. One candidate is the relationship of *co-reference*. That is, corresponding elements or terms in the two theories may refer to the very same entity, process or property in the world. As indicated by the quotes from Putnam and Laudan above, this is historically the most popular candidate amongst realists. The other candidate is the relationship of *structural correspondence*, whereby there is no point-to-point relationship between particular posits in the successive theories, but rather a shared conception of how the theoretical posits are understood to relate to each other. Both of these candidates, and related options, are discussed in far more detail in Chapter 3.

## 7. The optimistic induction

Given suitably clear definitions of “empirical success”, “essential” and “preserved”, DR can be construed as simply an empirical claim about the history of science. Whether or not a theory is empirically successful can be interpreted in such a way that it depends only on the empirical consequences of the theory and whether those consequences are actually observed (at some point in time). What parts of a theory are essential can be determined by careful examination of the logical relations between its postulates, and the relations between these and the results of empirical observation. And whether some elements of a theory are preserved in its successor can, as discussed in the previous section, be established by a comparison of the logical structures of the theories in question. These claims are obviously somewhat sketchy, and demand considerable filling out. For instance, one might have serious doubts about whether the logical relations between the postulates of a theory are really so transparent to the historian of science. A more adequate articulation of these claims will come in later chapters. For the moment, it is sufficient that there is nothing about any of these terms that is *intrinsically* dependent on some metaphysical notion.

Since DR is an empirical claim about the history of science, deciding on its truth or falsity requires some critical examination of the history at issue. Such critical examination is, of course, already a stock-in-trade of the existing scientific realism literature. The dominant approach in this literature to date is (speaking very broadly) “Popperian” (Popper, 1963/2002a; 1959/2002b), in that it is focused on putative counterexamples to DR. Anti-realists attempt to argue that there have been many episodes where the essential elements of empirically successful theories have *not* been preserved in successor theories. This, as argued above, is the substance of Laudan’s PMI. A reasonably comprehensive list of many of the putative counterexamples suggested so far in the literature can be found in Chapter 5.1. Realists are *also* focused on putative counterexamples to DR (as opposed to, say, confirming instances), albeit with the intention of showing they are not so problematic after all. This is Popperian precisely insofar as these putative counterexamples amount to “severe tests” for the realist’s hypothesis, and this hypothesis is “corroborated” to the extent that it passes these tests. This Popperian strategy could possibly be criticised in very general methodological terms, especially since this thesis partly addresses the very issue of theory confirmation (see Chapter 2). It will, however, be assumed that this strategy is broadly adequate. In the remainder of this section, the focus will be instead on how this strategy might be applied more consistently and rigorously and so amount cumulatively to an ‘optimistic induction’ in support of DR. Along the way, some arguments made by particular defenders of realism will also be criticised.

One of the major questions that concerned Popper is what should be concluded if counterexamples to a hypothesis do, in fact, emerge. Popper himself allowed for two options. Firstly, one could discard the hypothesis. Alternatively, one could modify the hypothesis or one of its auxiliary assumptions in such a way that it is no longer falsified. However, he cautioned that, if sufficiently many such *ad hoc* modifications are accumulated, the entire system of hypotheses will be rendered unfalsifiable and hence trivial. Some of the key auxiliary “hypotheses” associated with DR are the definitions of the terms “empirically successful”, “essential” and “preserved”. Consequently, there are three broad cases where inappropriate definitions of these terms could render DR trivial.

Firstly, the essential elements of theories could be picked out in such a way that they are always identical to those elements which we now happen to know were preserved in successor theories. Stanford (2003a, p. 914; 2006, pp. 166-168) and Newman (2010, pp. 416-417) argue that Psillos and Kitcher’s versions of partial realism fall into precisely this trap. As a pure exegetical matter, it is unclear whether this complaint is sustained; these authors can legitimately be read either way. The general point is that, for the thesis of DR to avoid triviality, “essential” should be defined in such a way that what parts of a theory are counted as essential does not depend upon this type of hindsight. This requirement, and the exegetical question, is discussed in much greater detail in Chapter 3.3.

The second sort of case is that where the terms “empirically successful” or “essential” are defined too narrowly. The consequence of this is that no actual examples satisfy them, and hence no counterexample is possible. This problem is not actually manifested in any version of partial realism advocated in the literature. The one area where it might be argued that partial realists succumb to it is when they rule out putative counterexamples as instances of “immature” science. In defining “mature” science, these authors do generally invoke a relatively strict criterion of empirical success, such as novel predictive success. This move is, however, not so extreme as to be illegitimate. The proposed definitions of success remain sufficiently wide to admit many cases and therefore many potential counterexamples. Moreover, and perhaps more importantly, these definitions are generally motivated by *priori* concerns about what sort of empirical accomplishments legitimate application of the NMA, rather than the goal of excluding problematic cases as such (see e.g. Zahar, 1973; Worrall, 2002; 2006). The notion of empirical success is discussed in much greater detail in Chapter 2.

The third case is that where the notion of theoretical continuity or preservation is defined too broadly. One set of scientific realists who have seemingly fallen into this trap are those who think that the important kind of continuity in science is *referential* continuity, and accept a causal account of reference (see e.g. Hardin & Rosenberg, 1982). Under such an account, the reference of a term is simply whichever actual entity the first person to coin the term was causally interacting with when she coined it. The problem with this sort of account, as pointed out by Laudan (1984), is that it seemingly makes it “too easy” for us to regard past theories as referring to entities posited by our existing empirically successful theories. Referential continuity approaches are discussed in more detail in Chapter 3.2.

Even a version of DR that avoids the various traps outlined above is not necessarily conclusively falsified by a single counterexample. For there is an additional auxiliary hypothesis involved, namely one which asserts that DR is to be understood as a universal claim, as opposed to a regularity which is merely satisfied most of the time. In the face of a single definitive counterexample, it is always possible to drop this assumption. Indeed, there is in fact no obvious *a priori* reason to construe DR as a perfectly general claim in the first place. This is true even for those whose commitment to it is motivated in the first instance by a commitment to (partial) scientific realism. The main argument for realism is the NMA, which rests on the intuition that we should disbelieve any explanation of an event which portrays it as simply a lucky accident (provided, of course, another explanation is available). And yet the proponent of this argument must accept that accidents do happen occasionally. The realist can therefore believe, without contradiction, both that empirical success gives a good reason to believe in the truth of a theory, and that successful theories will occasionally turn out to be false. The occasional counterexample is therefore not a falsifier to the realist’s general expectation of theoretical continuity.

Some of the recent work by Saatsi and Vickers is extremely insightful when placed in the context of the present discussion. These authors accept at least a few counterexamples to DR, including Kirchhoff’s diffraction theory (Saatsi & Vickers, 2010) and Bohr’s “old” quantum theory (Vickers, 2012). Focussing on Kirchhoff’s theory, they argue that certain false assumptions of Kirchhoff’s theory were essentially involved in deriving the accurate predictions of this theory. It is unimportant for present purposes whether this reading of the historical record is correct[2]. What is interesting is that they attempt to show that this counterexample does not pose a general threat to realism.

They do so by attempting to demonstrate that "*the world in this specific respect* is very kind to a human scientist who works her way upwards from the simplest assumptions" (Saatsi & Vickers, 2010, p. 43, emphasis in original). That is, they offer an account, from the modern perspective, of *why* the false assumptions nevertheless resulted in accurate predictions. Such an account potentially removes the threat to realism precisely because the NMA has as a premise that achieving empirical success with a false theory is highly unlikely. If it turns out that, under in certain circumstances, false assumptions are actually quite likely to result in empirical success, then this premise is violated. This certainly limits the applicability of the NMA. But it also eliminates the instances in question as potential counterexamples to the argument. The realist can still freely apply it in circumstances where prior investigation reveals that the crucial premise *does* hold.

Saatsi and Vickers do, however, reject the following general argument for not interpreting individual counterexamples as refutations of scientific realism:

> "One might assume that, given the vast literature on the realism debate, there cannot possibly be many other historical examples like Kirchhoff's hiding in the woodwork. Hence, the suggestion is that perhaps in this case nature really has been unusually kind to us ... and Kirchhoff's success from falsity is just a one-off." (*ibid.*, 2010, pp. 42-43)

This is very similar to an argument given above, which hinges on the claim that the NMA does admit of occasional exceptions. They do not reject it on any *a priori* grounds, however, so much as by questioning the claim that the literature on realism is as comprehensive as advertised. As Saatsi remarks in an earlier paper:

[2] Though note that Vickers (forthcoming), at least, now appears to concede that this example can be accounted for within the partial realist framework. See Chapter 3.9 for a further discussion of this case study.

> “Most realist commentaries on historical theory-shifts focus on just one or two oft-repeated cases of potential reference invariance and ontological discontinuity. Most commentators have taken as their starting point Laudan’s infamous list which, Laudan alleged, could be “extended ad nauseam”. It seems that most of the ensuing literature has focused on tackling the cases explicitly mentioned by Laudan, and very few case-studies have been sought beyond the list. ... *Prima facie*, the immense breadth of scientific enterprise over the past couple hundred years (say) bears promise of many more cases to be discussed in order to get a faithful overview of theory change in science.” (Saatsi, 2009)

Saatsi is certainly descriptively accurate on the state of the literature. Worrall states that “… some of Laudan’s examples remain to challenge the realist. Let’s concentrate on what seems to me (and to others) the sharpest such challenge: the ether of classical physics” (1989, p. 115). Psillos, also implicitly referring to Laudan’s work, promises that his discussion “… illustrates and strengthens the realist defence against the pessimistic induction by considering the two mature and genuinely successful theories alleged to have been characteristically false: the caloric theory of heat and nineteenth-century optical theories” (1999, p. xxiv). Kitcher (1993, pp. 140-149) and Chakravartty (Chakravartty, 2007, ch. 2) both also reference Laudan and yet again give the example of nineteenth-century theories of light pride of place in outlining their respective versions of partial realism.

Saatsi’s complaint is quite naturally cast in Popperian terms, and so can provide some additional guidance for the strategy of providing an “optimistic induction” in support of DR. The point is that it is difficult to avoid the suspicion that each of the authors cited above is more concerned to *illustrate* the particular version of partial realism he favours, rather than subjecting it to severe testing. After all, the basic treatment of the Fresnel example has already been established by Worrall (1989a) and Worrall traces his analysis back as far as Poincaré (1905/1952). In each author’s theory, the ether itself is discarded as inessential, and Fresnel’s wave equations for light are preserved, although the rationale differs slightly in each case. There is, of course, nothing wrong with using a neat, well-known example to illustrate the basic logical apparatus of some hypothesis. But to actually provide *evidence* for a hypothesis like DR, an investigator must seek out *additional* putative counterexamples.

Saatsi levels another, related, complaint in the same paper:

> “Like so many other problems in philosophy, the question of realism has been posed in extremely general terms at the outset, and considerations such as the Miracles argument have been put forward in the hope to secure realism about “all or most of mature science” in one fell swoop. I will now propose that the realism debate can progress by renouncing such extreme level of generality in its arguments and acknowledging the heterogeneity and the multifaceted nature of theoretical science, and by studying our epistemic commitments in a more piecemeal way.” (Saatsi, 2009)

In other words, a successful test of DR in the context of one scientific discipline provides only relatively weak evidence that it is generally true. This is particularly a concern in light of the preponderance of examples from physics in the literature. The picture in molecular biology, for instance, may be quite different. So anyone who aspires to a successful optimistic induction in favour of DR must not only examine a large *number* of putative counterexamples, but also examine such cases in various scientific disciplines.

The foregoing discussion in this section does not radically change the terms of the debate over scientific realism in the history of science. Scientific realists and anti-realists still will (and should) clash over the interpretation of putative episodes of theoretical discontinuity. This discussion does, however, suggest some conditions by which this debate might be carried out more rigorously. Firstly, care should be taken to ensure that the key terms in DR are not defined in such a way that the position is empirically vacuous (or unfalsifiable!). This means, for instance, not picking out what is “essential” in some older theory by reference to what we already know has been preserved in successor theories. Secondly, realists should make some effort to pick out *new* putative counterexamples and examine them in detail. This is because, providing a “severe test” for some hypothesis involves examining circumstances where the outcome of the test is not already known to be favourable. Thirdly, realists should make more effort to examine examples in more varied scientific contexts.

## 8. Deflationary realism and constructive empiricisim

I focus particularly on van Fraassen’s (1980) “constructive empiricism”, since this is probably the most currently influential brand of empiricism. Van Fraassen agrees with the semantic claim of scientific realism, that scientific theories ought to be understood as making literal claims about the world. However, he rejects the no-miracles argument and generally claims that we can never know one way or the other whether propositions concerning unobservable entities are true. The main aim of science is therefore, in his view, to account for the *observable* features of the world. To accept a theory is merely to believe that it is “empirically adequate”, i.e. accurately describes observable phenomena, past and present.

In van Fraassen’s introductory description of his position, he makes the following remark:

> “... [A]cceptance [of a theory] involves not only belief but a certain commitment. Even for those of us who are not working scientists, the acceptance involves a commitment to confront any future phenomena by means of the conceptual resources of this theory.” (*op. cit.,* p. 12)

I interpret “a commitment to confront any future phenomena by means of the conceptual resources of this theory” as a commitment to something like the continuity of science (or at least the bit of science in question). This seems to suggest that, under van Fraassen’s view, a scientist who accepts a given theory S will endorse some proposed successor theory L only if it preserves some essential elements of S that make this predecessor theory successful. It seems, in other words, that van Fraassen endorses DR!

This interpretation is at least complicated, however, by a remark which follows shortly after the passage quoted above:

> “This is a preliminary sketch of the *pragmatic* dimension of theory acceptance.” (*op. cit.,* p. 12, emphasis in original)

The notion of a “pragmatic” commitment can be interpreted in at least two different ways[3]. On the one hand, the pragmatic grounds for a commitment to theoretical continuity might emerge from facts about how scientists actually think. On the other hand, this commitment might emerge from principled arguments about how scientists *ought* to theorise. DR is compatible with only the latter interpretation. It is unclear from *The Scientific Image* which of these interpretations van Fraassen intends, and I shall not attempt to resolve the exegetical question here. For current purposes, it is sufficient that the latter interpretation, and thus DR, is compatible with constructive empiricism. In the remainder of this section, these different interpretations will be spelled out in somewhat more detail. It will then be demonstrated that the constructive empiricist can in principle accept the optimistic induction outlined in the previous section.

First, if only to provide a point of contrast, it is worth spelling out the interpretation under which van Fraassen’s pragmatic commitment emerges from a claim about human psychology. Suppose it is the case that a scientist’s acceptance of a theory involves a certain reconfiguration of her intellectual apparatus such that it would generally be easier for her to work with the concepts of that theory in future. If this is the case, a scientist may be well-advised to use the concepts of the theory she has already accepted in future theorising. It does not follow from this claim, however, that the conceptual resources of the current theory will be any more successful in confronting future phenomena than those of any other actual or

[3] My thanks to Anjan Chakravartty and Eleanor Knox for pointing this out to me.

hypothetical competitor theory. A different scientist, who has accepted a different theory, might be well-advised to use the concepts native to *that* theory. This reading of constructive empiricism suggests a world where a large variety of theories, each employing different concepts, will be empirically adequate for any given class of phenomena. Acceptance of a theory does not therefore give any *general* reason to prefer successor theories that are substantially continuous with it. Given acceptance of a theory by some group of scientists, however, they will quite legitimately favour such theories by a kind of intellectual 'inertia'.

Under the second reading of van Fraassen's claim, a commitment to the conceptual resources of a given theory is motivated by some argument suggesting that these conceptual resources will generally be more helpful than any others that are available. One form such an argument may take is a purely 'local' inductive inference – if the theory is currently more empirically successful than its competitors, then it might be expected that the conceptual resources of this theory would continue to give empirical success. Although this argument does not make any realist appeal to the truth of the theories concerned, it is still susceptible to the pessimistic meta-induction however. The aim of the PMI is precisely to show that we *cannot* in general expect the concepts of a successful theory to continue to produce success.

As argued previously in this chapter, however, the pessimistic induction might be replaced by an optimistic induction over the history of science if the focus is placed more narrowly on the "essential" elements of empirically successful theories. This optimistic induction gives us reason to believe that *some* of the conceptual resources of current theories are likely to underpin whatever theories happen to be empirically successful in the future. And, since scientists will generally want new theories that are empirically successful, a justified expectation that certain concepts of a current theory are likely to be preserved gives these scientists a very good *pragmatic* reason for using these concepts in trying to devise the theory's successor.

It is worth repeating what was noted at the beginning of the previous section, namely that DR is a purely empirical claim about the history of science. The optimistic induction in support of DR, moreover, makes no appeal to unobservable entities or any other metaphysically contentious issues. It simply aims to show that a general empirical claim is supported by a particular class of empirical evidence (i.e. case studies from the history of science). Thus neither DR nor the optimistic induction is incompatible with the principles that the constructive empiricist already accepts. Thus, if the optimistic induction is sound, the constructive empiricist should accept DR.

This is all rather abstract, however. One relatively concrete way we might make DR more palatable for the constructive empiricist is by thinking of it as a *methodological* generalisation. There are many of these, of the form “When scientists in situation A use methodology B, they tend to derive a successful hypothesis”. Some examples of methodologies that can be substituted for B include “Write down the Hamiltonian” or “Identify a protein interaction”. However, since DR is very broad in scope, more appropriate analogies may be along the lines of “Write an equation that expresses an important relation” or “Look for a mechanism”. DR can be expressed in these terms as “Be conservative (in a particular way)”. Each of these methodological rules is supported by some inductive base, that is, a series of occasions upon which acting in the recommended way did in fact yield good results. So DR is thus far on equal terms with any other rule (provided that the inductive base gestured towards in section 7 does, in fact, exist). Thus the constructive empiricist could, and perhaps should, accept DR.

## 9. The argument from continuity to truth

DR is, on its own, simply a ‘brute’ empirical regularity; a general statement to be tested against some class of evidence. But the realist could argue that it is simply *unscientific* to posit a brute regularity without making any attempt to explain it. And, moreover, she could argue that one should in general believe the *best* available explanation of the regularity observed. To forestall a possible objection, it should be emphasised that there is no obvious distinction in this respect between regularities in history as it does for those of, say, physical systems. For instance, if we accept IBE as a valid form of inference in general, and if the best explanation we have the regular cycle of financial 'boom and bust' is that people are disposed to accumulate debt when economic prospects are good, and to save when prospects are poor, then we have good reason to believe that this mechanism in fact operates. Supposing now that there is substantial theoretical continuity in the history of science, and that the best explanation for this is that empirically successful theories are usually at least approximately true then we ought, by the same form of reasoning, to believe that the theories in question are approximately true.

It is worth pointing out that this is argument is distinct from the NMA, although both are instances of IBE (see the discussion in section 2). Whereas the "retail" version of the NMA makes the inference from the success of theories considered individually to their (approximate) truth, the proposed argument would infer from the continuity of theories in some historical sequence to the truth of all the theories in this sequence. This has some intuitive appeal, as it seems very implausible that there would be substantial continuity between two theories chosen at random. It is more plausible that the theories are similar to each because both accurately describe some independent external state of affairs. Call this the "argument from continuity to truth" (ACT). It can be formalised as follows:

ACT

P1: There is substantial theoretical continuity between theories S and L, both of which are empirically successful

P2: We should believe the best ("loveliest") of the available explanations for the theoretical continuity between S and L

P3: There are exactly two available explanations for this theoretical continuity: (a) that S and L are both (approximately and/or partially) true in respect of those features where there is continuity between them; and (b) that the continuity is merely result of an unlikely coincidence (a “miracle”)

P4: Positing the occurrence of an unlikely coincidence is never the best explanation if any other options are available

C: *Therefore*, we should believe that both S and L are (approximately and/or partially) true in respect of those features where there is continuity between them.

As with the NMA, no substantive points advocated in this thesis depend on the soundness of the ACT. However, since it is a novel argument, it is worth providing some critical analysis of it. My approach is to modify responses to the NMA so that they stand as plausible objections against the ACT. All the general arguments against the validity of IBE as a mode of reasoning discussed in section 2 apply here without modification, and so these will not be discussed these.

The first argument to consider against the ACT is a version of the base-rate argument. Suppose there is an older and a newer theory in a given domain, both of which have been extremely successful. Moreover, it is clear that there are substantially similarities between them. There are two scenarios where this outcome may arise: the theories are both (approximately) true; and the theories are both false *in the same way*. As before, the simple fact that both theories are successful increases the probability that they are true, but we cannot conclude that they are likely to be true without knowing what proportion of the class is composed of cases involving false theories. One possible rejoinder to this is to point out that, in contrast to the case with the (retail) NMA, there are now *two* theories involved. If a false theory is unlikely to be successful, surely it is *astronomically* unlikely that two theories which are false and similar to each other could both be successful?

This rejoinder is considerably weakened, however, by conjoining the base-rate argument with a selectionist argument. As described above, substantial continuity between successful scientific theories can be explained by postulating that these theories are latching onto some stable unobservable features of the world. But it can also be explained by postulating that these theories were *designed* to be substantially similar in certain respects. It has generally been the case that the scientists who designed a theory to account for some domain of phenomena were aware of (or, indeed, intimately familiar with) the existing theories in that domain. It seems quite plausible, then, that they might deliberately incorporate some of the features of the old theory into the new. There are several motivations that could underlie this conservatism. For instance, if some effort has been made to render tractable a set of mathematical equations found in the old theory, there may be a strong incentive to preserve the same formalism in the new theory. It may even be the case (in fact, I suspect that it frequently *is* the case) that the scientists involved tacitly endorse some version of DR and are therefore motivated specifically to be conservative.

It should be clear how the base-rate and selectionist arguments together constitute a difficult problem for the ACT. The first suggests the possibility that there are explanations for theoretical continuity between two successful theories other than their truth. And the selection argument picks out a particular explanation that is relatively plausible. It is *prima facie* highly unlikely that two false theories could be both empirically successful and very similar to each other by lucky coincidence; this, surely, would be a miracle. But all that the current scenario requires is that a single false theory be successful. Selection then ensures that the other theory is similar to it and, because they will be false “in the same way”, that this other theory will also be empirically successful.

For the main thesis of this paper, the chapter and the responses to it are by-the-by, though they may be of some interest to the partial realist. Before concluding this section, however, it is worth pointing out that the argument around selection also poses a potential difficulty for DR itself. Recall that the argument for DR is a form of “optimistic induction”. Positive instances for this induction are those cases where the elements essential for the predictive success of a given theory are also found in subsequent successful theories in the same domain. If, in a large proportion of these cases, the scientists involved in constructing the later theories are specifically trying to achieve theoretical continuity, then this induction becomes considerably more limited in scope. One cannot legitimately argue from such an inductive base in favour of the general claim that there will always (or typically) be theoretical continuity in science, but only that there will be such continuity *where scientists are trying to achieve it*. But one of the major motivations for arguing in favour of DR is that this thesis provides a justification for methodological conservatism. And such a circumscribed version of the thesis, although it justifies the inference that conservative theories are unlikely to be a waste of time, does not weigh against non-conservative theorising. The issue of selection, if indeed it is widespread in the history of science, may therefore render DR considerably less interesting and useful.

For the defender of DR, there are at least two possible responses to this problem. The first response is to notice that the selection argument inherits from the base rate argument the presupposition that many different scientific theories could potentially be successful in a given scientific domain. The argument fails if this presupposition is defeated. Putting this another way, the fact that a scientist came to the new theory by working from the old theory would not be a substantial blow against either DR or the ACT if the new theory was one of only a very small number that had any chance of being successful anyway. It is, however, difficult to see how to make convincing the claim that the ‘population’ of potentially successful theories is small. The most prominent *a priori* argument that addresses the question is the argument for the underdetermination of theory by evidence, which purports to demonstrate precisely the opposite conclusion. The prospects of *a posteriori* arguments would also appear to be poor. Suppose even that in all actual cases only a small number of theories have been proposed as potential successors to the existing successful theory in some domain, and of these a large proportion have been substantially continuous with the older theory. This evidence cannot defeat the presupposition that there is a large population of *logically possible* theories that could potentially have been successful. Moreover, appealing to this evidence simply begs the question, as a proponent of the selection argument would presumably argue that a large proportion of actually proposed theories match our criterion precisely because of selection.

The second response to the problem draws on Lipton's decisive response to the original version of selectionist argument against the NMA, both of which were discussed in section 3 of this chapter. Recall that the major weakness with the original version of the selectionist argument is that it is unable to account for cases where a theory predicts qualitatively novel phenomena. This is for the very simple reason that selection is intrinsically a backward-looking procedure – selection cannot account for a theory's empirically adequacy in respect of phenomena that were not known when the theory was accepted by the relevant scientists. Similarly, to defeat the selectionist argument against the ACT, cases must be highlighted where these is theoretical continuity that cannot be accounted for by selection. These cases will never be as watertight as those of novel empirical prediction. The proposition that a scientist selected a theory specifically for its ability to explain particular phenomena can be ruled out definitively in cases where these phenomena have not yet been observed. It is, on the other hand, impossible to rule out entirely the proposition that a scientist has crafted a successor theory in such a way that it preserves elements of an earlier theory. Nevertheless, there are cases where this is historically implausible as an account of how a theory was developed.

To illustrate this, I will briefly examine a historical example where DR is apparently sustained, but there is no evidence of selection. The example is Maxwell's reformulation of the wave theory of light (the historical details are provided by Mahon (2003, ch. 7). Around 1819, Fresnel defended a wave theory of light and, motivated by the results of polarization experiments, postulated that it consists of a transverse wave propagating in an elastic solid "ether" which pervades space (see Worrall, 1989b; Fresnel's theory is also discussed in more detail in Chapter 2.3). Around 1862, Maxwell was concerned with a problem that is, on the face of it, entirely different to Fresnel's, namely that of accounting for electrical and magnetic phenomena. At this time, these were already understood to be related, for example by the phenomenon of electromagnetic induction. Maxwell constructed a mechanical model whereby space is filled with spinning “cells” or “vortices”. Magnetic lines of force are constituted by these cells lining up along their spin axes, and electric currents occur in a conducting material when tiny electric charges are passed from one cell to the next. To account for the electrical properties of insulating materials, Maxwell postulated that the cells are elastic and that an electric field is constituted by the distortion of these cells across a region of space. This model satisfactorily accounted for the electromagnetic phenomena of immediate interest. But a secondary consequence of this model is that the postulated medium has the right mechanical properties (notably elasticity) to fulfil the role of Fresnel’s ether in sustaining the propagation of transverse waves. And, when Maxwell carried out the relevant calculation with empirically measured values for the electromagnetic and electrostatic units of charge, he predicted the occurrence of waves with very nearly the same velocity as the measured value for light. The electromagnetic theory of light which grew out of this insight rapidly became dominant in optics. And this theory had substantial continuities with Fresnel’s earlier theory, not least the postulate that light is constituted by a transverse wave. Yet it is difficult to argue, in light of the history sketched above, that the electromagnetic theory was selected precisely because of these continuities. The new theory emerged from an investigation of electricity and magnetism, both of which are *prima facie* quite distinct from optical phenomena.

To generalise from this case study, it seems that far greater emphasis should be placed on those cases where scientists have ‘unintentionally’ adopted, or even been ‘forced into’ using, the theoretical concepts of the predecessor theory. This more limited inductive base might then be used to construct an argument for the stronger form of DR. The appeal of this response is that it apparently circumvents the problem of selection in a single stroke. It is not at all clear, however, how typical the Maxwell case is. The strategy of limiting our inductive base to cases where selection is absent may therefore be unfeasible. I do, however, take this to be an open empirical question for the historian of science. The issue of selection will not be addressed in the majority of the case studies examined in the remainder of this thesis, where the aim is merely to assess the simple version of DR against various putative counterexamples. It will, however, be discussed briefly in connection with the phlogiston theory of chemistry, in Chapter 5.6.

## **10. Chapter summary**

In this chapter, I have critically analysed the most prominent arguments respectively for and against scientific realism, namely the NMA and the PMI. Given the conclusion that the PMI is sound, what I have termed “naive realism” is untenable. However, one or other version of “partial realism” may avoid this problem by circumventing the PMI’s claim that “radical discontinuity” is characteristic of theory change in science.

I have proposed “deflationary realism” as offering a new perspective on the scientific realism debate. DR is, in essence, a *direct* commitment to theoretical continuity in science, one that can be defended or attacked on its own terms. The advantage of DR is that it is simultaneously weak enough to avoid irresolvable metaphysical disputes yet strong enough to have interesting consequences. It is weak because it claims simply that there is an empirical regularity in the history of science, namely a certain sort of continuity between predictively successful theories that have existed in the same scientific domain. It is not, however, committed to any particular explanation of this regularity. DR is strong because, if this regularity can be held to obtain between our existing best theories and their putative successors, it justifies a methodological commitment to conservatism in scientific theorising. Moreover, because DR is entailed by PR, a refutation of the former is also a refutation of the latter.

DR can be supported by a sort of “optimistic induction” over the history of science. I have claimed that the most promising approach to such an induction is essentially falsificationist. That is, both realists and anti-realists are well-advised to focus on those historical cases that *prima facie* fail to satisfy DR, though realists would typically do so in the hope that these cases turn out not to be counterexamples after all. For the falsificationist strategy to be effective, however, two broad conditions must be satisfied. Firstly, DR must be formulated in way that is falsifiable. So, for instance, the criterion for which theories are empirically successful should not be so narrow that no actual cases satisfy it. Secondly, DR must be subjected to severe tests. That is, we should not focus only on particular cases or scientific sub-disciplines where there is already good reason to believe DR is satisfied. We should also avoid cases where there is evidence that the scientists involved specifically intended the successor theory to be substantially similar to its predecessor.

I have argued that the constructive empiricist could in principle accept DR and indeed, under a certain reading of van Fraassen, is already committed to it. Finally, I examined an argument (the ACT) that the partial realist might use to resist the deflationary tendencies of DR. This argument maintains that a historical pattern of continuity between scientific theories demands an explanation, and that the best explanation is the approximate truth of the theories concerned. Although I am officially neutral on metaphysical questions, and so do not wish to deny this argument to realists, I have nevertheless shown that the anti-realist has some very convincing responses to it. DR therefore stands as a defensible position in its own right.

In the following two chapters, I will focus on disambiguating some of the vague key terms found in the definition of partial realism and thus that of deflationary realism, namely “empirical success” and “essential”. It is worth noting that I will attempt to give definitions of these terms from a partial realist perspective. For instance, I will offer a definition of empirical success such that a theory satisfying this definition might be regarded as likely to be (approximately) true under the NMA. Adopting this perspective is not strictly necessary if our aim is simply to articulate deflationary realism as a standalone position, one that may be defended by an optimistic induction. For this purpose, any definitions that ensure the overall falsifiability of the position will be sufficient, as discussed in section 7 above. Nevertheless, it is worthwhile, at least in the first instance, retaining the connection between our present discussion and the larger realist debate. This ensures that any putative refutation of DR is also a refutation of PR. It also means that, if either the NMA or the ACT are judged to be viable arguments, that the version of DR defended here can simply be ‘re-inflated’ to yield a corresponding version of PR.

The discussion in the following two chapters is, in sum, intended to result in defensible versions of both DR and PR. These first three chapters, then, comprise the ‘core’ position of the present thesis. In subsequent chapters, I then turn to addressing some of the consequences of the core position, and examine how well it accounts for particular putative counterexamples. In Chapter 4, I will address a key consequence of partial realism, namely the particular view of scientific rationality it entails. I will examine this view in the context of the long-running philosophical literature surrounding “scientific revolutions”. Finally, in Chapter 5, I will examine in detail some case studies that potentially cast doubt on deflationary, and thus partial, realism.

# Chapter 2. Empirical success and theory confirmation

## 1. Chapter overview

The aim of this chapter is to give a defensible account of when a theory should be judged “empirically successful”. This is particularly relevant to a full understanding of partial realism and deflationary realism, since the term appears in each of their definitions. The major criterion I will adopt in assessing any proposed account of empirical success is that, supposing the NMA a valid form of argument, it should be plausible to infer the truth of a theory which enjoys empirical success under the proposed account.

I begin, in section 2, with a brief general overview of the idea that a theory is more confirmed by novel prediction than by accommodation of empirical results. In this discussion, a distinction is made between logical and historical theories of confirmation. The most natural interpretations of the notion of novel prediction are historical in nature, and these are discussed in sections 3 - 5, culminating in a simple version of the “use-novelty” (UN) account. As these accounts are all subject to criticism under the criterion state above, however, more logical accounts of empirical success will come to be favoured.

The focus of sections 6 and 7 is on accounts of confirmation which view novel prediction as simply a special case of a more general theory of confirmation. The “severe testing” account, and criticisms of it, are discussed in section 6. Some accounts that regard the ability of a theory to unify disparate empirical results as the primary theoretical virtue are discussed in section 7. These accounts regard predictive success as merely a secondary virtue, and hence are grouped under the heading of “weak predictivism” It is argued that these accounts, unlike the other accounts survey in the chapter, are broadly adequate to the NMA.

Consequently, a specific “unification view” (UV) is developed in section 8. This is developed from Worrall’s writing on the UN account of predictive success, though ultimately it is a version of weak predictivism. It asserts that a theory is confirmed (has “unifying power”) to the extent that it entails more verified empirical results than are required to construct it.

Sections 9 and 10 are devoted to addressing some problems with the initial version of the UV proposed above, and consequently suggestion some modifications to it. Section 9 focuses on a key ambiguity in the idea of “constructing” a theory, and it is ultimately argued that a much more expansive definition of this term than that found in section 8 is required. Although the initial definition of the UV emphasises the number of empirical results *entailed* by a theory, the thesis in section 10 is that deductive entailment of results is neither necessary nor sufficient for a theory to count as unifying. The chapter is summarised in section 11.

## **2. Predictivism**

For the purposes of the NMA, “empirical success” is frequently cashed out by reference to the notion of “novel prediction”. This fits well with the common view amongst philosophers of science that, *ceteris paribus*, a theory is better confirmed by the prediction of some novel phenomenon than it would be by the “accommodation” of the same phenomenon after it has already been observed. Call this view, following Maher (1988; 1990) “predictivism”.

Predictivism has a long provenance in science and philosophy of science. Musgrave (1974) cites the example of Thales (circa 600 BCE) being dubbed one of the “Seven Sages” on the strength of being able to predict an eclipse. However, the debate between predictivism and anti-predictivism in something resembling its present form first takes place between Whewell and Mill. Whewell writes:

> “The hypothesis which we accept ought to explain phenomena which we have observed. But they ought to do more than this: our hypotheses ought to *foretel* [sic] phenomena which have not yet been observed... Because the rule prevails, it includes all cases; and will determine them all, if we can only

> calculate its real consequences. Hence it will predict the results of new combinations... And that it does this with certainty and correctness, is one mode in which the hypothesis is to be verified as right and useful." (Whewell, 1840/1847, pp. 62-63, emphasis in original)

And later, more explicitly, he writes:

> "Such a coincidence of untried facts with speculative assertions cannot be the work of chance, but implies some large portion of truth in the principles on which the reasoning is founded... The prediction of results, even of the same kind as those which have been observed, in new cases, is a proof of real success in our inductive processes" (Whewell, 1858/1968, p. 152)

Notice here that Whewell, in claiming that "such a coincidence ... cannot be the work of chance" is appealing to the very same intuition underlying the NMA. In this chapter I shall, following Whewell, adopt the realist assumption that the empirical success of a theory can, *via* the NMA, be good grounds for believing that it is true. Given this assumption, our task is to find an account of novel prediction, or empirical success more generally, which suffices to render empirically successful theories well-confirmed under the NMA. The NMA, in other words, gives us a basic condition of adequacy for any account of empirical success.

Before diving into detailed analysis of the various accounts on offer, it is worth making a distinction, following Musgrave (1974), between logical and historical theories of confirmation. Under a logical approach, the degree to which a theory is confirmed by a piece of evidence depends entirely on the logical relationship between the two. In the simplest case, the theory deductively entails the evidence. Under a historical approach, the degree of confirmation depends at least partially on details of the historical situation in which the theory was proposed. It may therefore be relevant whether the theory was proposed before or after a particular observation was made, and/or what phenomena the theorist(s) who proposed the theory intended to account for.

## 3. Temporal novelty

Predictivism can be thought of as according special epistemic weight to *unexpectedness*. Recall that the NMA posits two potential explanations for the empirical success of a particular theory: either the theory is true, or a lucky coincidence ("miracle") has occurred. And (provided no other explanations are tenable), the explanation of truth is favoured precisely to the extent that a coincidence is unlikely. If a theory predicts some phenomenon which is already considered likely to occur on the basis of background knowledge, however, the agreement between theory and phenomenon is not at all unlikely.

This intuitive argument can be neatly formalised with reference to the Bayesian updating formula. This asserts that the correct assignment of subjective probability to a hypothesis (H) upon the observation of some new piece of evidence (E) is given by the conditional probability of the hypothesis on the evidence, as assessed *before* the observation. This conditional probability, in turn, is given by Bayes' theorem.

$$p_{new}(H) = p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}$$

Where:

- $P_{new}$(H) is the probability of the hypothesis *after* updating on the new evidence
- p(H|E) is the probability of the hypothesis given some piece of empirical evidence (the 'posterior probability')
- p (E|H) is the probability of the evidence assuming the hypothesis is true
- p (H) is the probability assigned to the hypothesis without assuming that this evidence has been uncovered (the 'prior probability')
- p (E) is the probability of the evidence without assuming that the hypothesis is true (i.e. the 'background' probability of the evidence)

In the simplest case, the evidence is entailed by the hypothesis, so the term p(E|H) is unity and so can be eliminated from the right-hand-side of the equation. The prior probability p(H), i.e. the degree to which we believe the hypothesis before encountering the evidence, is obviously not affected by the evidence itself, and so can be treated as a constant. In this case the degree to which we change our belief in the hypothesis – i.e. the ratio between p(H|E) and p(H) – depends only on p(E). The *less* likely the evidence without the assumption that H is true, the *higher* the posterior probability assigned to the hypothesis if E is in fact observed. That is, the more unexpected the evidence that the evidence entailed by the hypothesis is on background knowledge, the more that this evidence serves to confirm this hypothesis.

This brings us to the temporal account of novelty prediction. This account defines novelty simply as a phenomenon hitherto unknown to science. A theory makes a successful novel prediction, therefore, simply in case it entails (or renders probable) some phenomenon that has not been observed before, but which is indeed observed when the appropriate experiment or systematic observation is conducted. From the perspective of natural language, this account certainly gives the most obvious interpretation of “novelty”. One thing that should be emphasised, however, is that this account accords far greater importance to the observation of *qualitatively* new phenomena, as opposed to additional observations of phenomena which the theory is already known to account for.

A paradigm case of novel prediction – the Fresnel-Poisson ‘white spot’ (Worrall, 1989a; 1989b) – illustrates the intuitive appeal of the temporal view. In the early nineteenth-century, the science of optics was divided between proponents of the particle and wave theories of light. Since interference is a phenomenon found only in waves, the most compelling point in favour of the wave theory was the existence of diffraction experiments, such as the two-slit experiment carried out by Young in 1803, which appeared to show interference effects in light. In 1818, Fresnel submitted a treatise for a prize competition announced by the French Academy on the subject of diffraction phenomena. In this work, he introduced a version of the wave theory which made mathematically precise the concept of interference and so produced quantitatively precise predictions for the results of diffraction experiments. He could predict, for instance, the banding pattern around the edge of the shadow cast by a straightedge from a point source. Poisson, one of the judges on the prize committee, and a vociferous opponent of the wave theory, determined that Fresnel's theory predicted that the shadow cast by an opaque circular disc illuminated by a point light source would have at its centre a region that was as intensely illuminated as it would be if there was no disc present (a "white spot"). This consequence, while intuitively implausible, was nevertheless confirmed experimentally by Arago, another member of the committee. So what was intended as a *reductio* of the theory emerged as a striking piece of evidence in its favour. And, claims the temporal account, because this observation was so unexpected, it confirmed the theory more than the successful treatment of known diffraction phenomena.

Intuitive as this is, there are however several problems with the temporal account. The first stems from the claim that to count as "novel", an observation must be hitherto not "known to science". But this is extremely ambiguous. As Musgrave puts it:

> "Does a statement become 'known to science' when some scientist first thinks of it? – or when he first writes it down? – or when it is first published? – or perhaps when it is abstracted in a leading journal?" (Musgrave, 1974), pp. 8-9)

He goes on, however, to dismiss the problem:

> "Often it will not matter how we answer, for a very loose dating will suffice. And where it does matter, historians are quite used to settling such questions." (*ibid.,* p. 9)

In other words, it will usually be possible to establish at least whether prediction or observation came first, although not necessarily by how much time.

This response, however, seems to miss the point. The problem, surely, is that, even where the historical facts of the matter are settled, it is sometimes unclear whether some phenomenon should be *counted* as known to science for the purposes of deciding whether a theory is confirmed. This point can be made more sharply by considering a further wrinkle in the case of the Fresnel-Poisson white spot. As a matter of historical fact, this phenomenon *was* observed about a century prior to Fresnel's submission to the French Academy, and these observations were reported in widely-circulated scientific journals (Delisle, 1715; Maraldi, 1723; Worrall, 1989b). It just so happens that none of the prize committee, or Fresnel himself, was aware of these results. The question for the advocate of the temporal account, then, is whether this new piece of information should alter the extent to which we view Fresnel's theory as confirmed.

This question poses a dilemma for a proponent of the temporal view. On the one horn of the dilemma, the fact that the white spot had been observed prior to Fresnel's articulation of his theory could lead us to regard the theory as less confirmed. This seems like a strange conclusion, however. What if, for instance, instead of being recorded in French scientific journals in the eighteenth century, the Fresnel-Poisson spot had been observed by Aztec astronomers in the sixteenth century, shortly before they all died of diseases introduced by Europeans? It would surely be absurd to claim that Fresnel's theory should be considered less confirmed if an experiment that was so temporally, culturally and geographically distant from his own work suddenly came to light. Intuitively, there must surely be *some* causal connection between the prior observation and the theory for the former to count against the latter's predictive powers. A strict interpretation of the temporal account therefore seems unsatisfactory.

On the other horn of the dilemma, we could accept that the degree of confirmation of the theory is not at all diminished by the earlier observation of the white spot.

Taken more broadly, this interpretation seems to suggest that the degree of confirmation depends on whether the predicted phenomenon was known to the actual scientist(s) who made the prediction, as opposed to “science” generally. This is perhaps the most sensible interpretation of the temporal account in any case. It certainly fits very naturally with the Bayesian framework discussed above, since the latter is most meaningfully applied to understanding belief revision in a *particular* agent, albeit often with many idealisations. And, as shown above, it can easily rationalise a scientist’s regarding as confirmed theories which predict consequences that are surprising *to her*. But this interpretation now seems *too historical* to be plausible as a general account of theoretical confirmation. Why should *we* regard a theory as better-confirmed just because it made predictions that were surprising to Fresnel, Poisson, *et al.*? Or, more formally, there does not appear to be an obvious logical connection between the subjective degree of confirmation that an individual (let us assume rationally) attributes to a theory from her own epistemic perspective, and the degree to which later scientists or philosophers of science should regard that theory as confirmed. This distinction between a particular scientist’s subjective reasons for regarding a theory as confirmed and some more ‘objective’ account will become more prominent in later sections, and so will be discussed in more detail there.

Another major objection to the temporal account is due to Worrall (1989b, p. 144), who argues that, *as a matter of historical fact*, temporal novelty did not play a very significant role in scientists’ decision-making around the white spot case. Worrall notes that the judges of the French Academy prize committee paid comparatively little attention to the novel phenomenon of the white spot. Instead, they focused the majority of their attention on Fresnel’s successful treatment of already known straight-edge diffraction phenomena. Moreover, Poisson realised that Fresnel’s theory would also give predictions for the case of a disc with a circular aperture in the middle illuminated from a point source. Yet no member of the prize committee actually bothered even to carry out the detailed calculation! It was left to Fresnel, after he had been awarded the prize, to perform the calculation and confirm it experimentally. So the prize committee did not consider it worth their while to follow up on the prospect of a second, possibly equally counterintuitive, prediction. These two examples suggest that the cream of the French physics establishment did not consider temporal novelty a matter of great importance in deciding whether or not to accept a new theory.

This particular historical example illustrates a more general problem for the temporal account. This is effectively articulated by Glymour (1980) in his discussion of the problem of "old evidence" for Bayesian confirmation theory. If some piece of evidence is widely accepted as a piece of background knowledge, the prior probability of the evidence in the Bayesian formula is unity. The posterior probability of the hypothesis is then just the same as the prior probability. So not only do temporally novel observations not confirm as much as novel predictions; they don't provide any confirmation at all! But this is manifestly contrary to actual scientific reasoning and practice (hence the *problem* of old evidence). When a promising new theory or hypothesis is introduced, at least part of the reason scientists typically cite for believing it is that it adequately explains existing phenomena.

There are some clear examples of theories being accepted on the basis of "old evidence". Copernicus' heliocentric model of the universe, for instance, was largely supported by astronomical data that had been recorded over a period of centuries before he did his crucial work. Glymour illustrates his own argument by reference to Einstein's General Theory of Relativity (GTR). One of the major pieces of evidence cited in favour of GTR at the time it was proposed is that it deals satisfactorily with the anomalous (under the Newtonian theory) precession in the perihelion of the planet Mercury. This phenomenon, however, was first observed in the mid-nineteenth century – more than sixty years before the publication of GTR. Finally, the problem of old evidence is also manifest in the Fresnel example – both the Bayesian view and the simple intuitive appeal to 'unexpectedness' fail to account for the judges' endorsement of Fresnel's theory on the basis of its successful treatment of *known* diffraction phenomena. It is reasonable to conclude, then, that the temporal account fails to recover the actual reasoning of scientists.

It is worth making explicit the intuition underlying all the objections raised against the temporal account above (variants of the following argument are made by various authors, including for instance Musgrave (1974) and Leplin (1997, esp. ch. 2). This intuition is that it seems entirely *arbitrary* whether or not a particular phenomenon is observed before or after the postulation of some theory. The degree to which a piece of evidence supports a theory should depend on some sort of *logical or epistemic* relationship between the two. Granted, it is sometimes the case that scientists only think to perform an experiment because they wish to test a theory which predicts a particular result from it. It is also certainly true that scientists can only begin to use a piece of evidence to argue in favour of a theory after that evidence has become known. But these relationships are, as it were, 'accidental'. A difference in the temporal order between the appearance of the evidence and that of the theory lacks any *intrinsic* epistemic import. When scientists and philosophers of science judge that an observational result which is well-known can nevertheless provide support to some newly-proposed theory, they perceive a relationship between the two which *is* epistemically significant. In the follow sections, I will therefore survey some accounts of novel prediction which purport to establish this type of epistemic significance.

## 4. Theoretical novelty

One possible way of providing a principled logical or epistemic distinction between novel and non-novel observational results is to emphasise novelty at the time a theory is proposed relative to the existing *theoretical* background, rather than relative to the *phenomena* that are already known. One proponent of such an account is Musgrave:

> "The basic question is not so much 'Does evidence *e* confirm hypothesis *h*?' but rather 'Does evidence *e* support $h_1$ more than $h_2$?'. This suggests that in assessing the evidential support of a new theory we should compare it, not with 'background knowledge' in general, but with the old theory which it

> challenges... According to this view, a new theory is independently testable (or predicts a ‘novel fact’) if it predicts something which is not also predicted by its background theory.” (Musgrave, 1974, pp. 15-16)

Leplin advocates a similar position. He states two conditions for “the prediction of an observational result O to be novel for a theory T”, namely:

> “Independence Condition: There is a minimally adequate reconstruction of the reasoning leading to T that does not cite any qualitative generalization of O.
>
> Uniqueness Condition: There is some qualitative generalization of O that T explains and predicts, and of which, at the time that T first does so, no alternative theory provides a viable reason to expect instances.” (Leplin, 1997, p. 77)

By the “qualitative generalization of O”, Leplin means the general phenomenon of which particular observational result O is an instance. The Independence Condition requires that the observational result in question not be essentially involved in the construction of the theory which entails it. This is very similar to the use-novelty condition which, since it is the focus of the following section, will not be examined any further here.

Considering the Uniqueness Condition by itself, however, gives a view that is very similar to Musgrave’s. Both views regard the prediction of an observational result by a theory as novel only if it predicts some phenomenon that all existing rival theories either forbid, or are silent upon. There are cases where this ‘theoretical novelty’ view coincides with temporal novelty. However, it can also be the case that phenomena which are not predicted by the old theory are already known, and are now predicted by the new theory. The new theory is confirmed by this ‘old evidence’ simply because it predicts them while its rival forbids them. The Musgrave/Leplin account is therefore more successful than the temporal novelty approach at recovering the judgements of the scientists involved in episodes of theory change.

This account does not satisfy the basic adequacy condition for an account of empirical success, however, because theories that are successful under this account will not generally be well-confirmed under the NMA. This is for the simple reason that this approach also introduces an irreducible element of historical contingency. With the temporal novelty approach, the degree to which a theory is confirmed by an observational result depends on the epistemically irrelevant question of whether the result came before or after the theory was proposed. Under the theoretical novelty account, the degree to which an observational result confirms a theory depends on what other rival theories have been proposed at the first moment when the empirical result is known. And this is surely also a result of historical 'accident' to some degree.

In response to this, Musgrave and Leplin might well argue that *of course* the degree to which we believe a theory ought to depend on what other theories are present. This response is accurate, but beside the point.  Recall that we want criteria for which cases of predictive success give good reason to believe that the theory (or parts thereof) responsible for these successes is (are) *true*[4]. These criteria must be stronger than those for cases where a piece of evidence merely gives reasons to *prefer* a theory over some competitor, or to accept it *pragmatically*. Having reasons to regard a theory as true is, of course, also to have reasons to prefer it over a competitor and to accept it pragmatically. But the converse does not follow. It is entirely sensible to accept, and scientists frequently do accept, theories on the grounds of empirical adequacy, mathematical tractability, etc. without forming a commitment one way or another regarding their truth.

[4] Our account of novel prediction could, of course, also allow for lesser degrees of confirmation. But it must allow such a strong degree of confirmation in at least some cases to be compatible with the NMA.

It is not at all clear how one could use the theoretical novelty account to argue for the truth of a theory under the NMA. It is, after all, possible for a theory to shine in comparison to its competitors simply because those competitors are grossly inadequate. Ptolemy’s model of the universe, for instance, is highly confirmed relative to a cosmology in which the earth is flat and the planets are capricious gods. More generally, the opponent of the NMA can always happily concede that some theories are *more* empirically successful than others, while maintaining that *none of them* is successful enough to warrant inference to its truth. To avoid this response, the NMA therefore demands an *absolute*, rather than a merely *comparative*, account of empirical success.

Such an absolute account cannot depend on any of the contingent historical factors that give rise to the problems described above. The simplest, though not necessarily the only, way to avoid such dependence is to require that any proposed account characterise a two-place relation between the theory and the phenomena it predicts. Such a relation, perforce, makes no reference to the particular times at which the phenomenon was discovered or the theory was first proposed, to the presence or absence of competitor theories, or even to the (non-) existence of other phenomena. Interestingly, Leplin, in an exposition of his account earlier in the same book cited above, advocates a “relational requirement” for any account of novel prediction, which states that “Novelty is a binary relation between [an empirical result] and [a theory]” (*op. cit.*, p. 63). As pointed out by Ladyman, however, the independence and uniqueness conditions together entail that “a result being novel for a theory seems to be a relation between a theory, a result, the provenance of the theory, and all the other theories around at the time since the latter are required not to offer explanations of the result” (Ladyman, 1999 p. 183). So Leplin’s fully-developed account of novel prediction is clearly inconsistent with a desideratum he himself proposed! In any case, for the reasons cited above, although Leplin’s particular view of novel prediction is unsatisfactory, his relation requirement, or at least the motivation behind it, is well-taken.

If confirmation is a binary relation between observational result and theory, then the inference from the result to the truth of the theory will remain intact whatever new theories emerge. It is also is extremely likely to remain (at least approximately and/or partially) intact whatever new *evidence* emerges[5]. At first blush, these consequences would seem to provide a kind of *reductio* argument against any account of novel prediction satisfying the requirement that confirmation be a binary relation. Surely all theories, no matter how successful, will very probably be totally overturned by later scientific activity! But this just a restatement of the conclusion of the PMI, and is just what the NMA denies. Some reasons for believing a theory – including the fact that it predicts novel phenomena, according to the realist authors under discussion – are so decisive that it should be viewed as extremely likely that the theory (or at least the parts of it essential for its predictive success) will *not* be overturned. It would, so to be speak, be a "miracle" if the theory later turned out to be false. This is not, of course, to say that the NMA is necessarily successful. Even under the more careful partial realist reading, it may well be the case that theories which enjoy novel predictive success turn out to be radically false. The narrower point is that is that any account of confirmation that is adequate for the NMA, whether or not the NMA is successful, must satisfy Leplin's relational requirement and so be insensitive to later scientific developments in the manner described.

## **5. Use-novelty**

The so-called "heuristic" account of novel prediction is characterised by Zahar as follows:

> "A fact will be considered novel with respect to a given hypothesis if it did not belong to the problem-situation which governed the construction of the hypothesis" (Zahar, 1973, p. 103).

Worrall, along similar lines, argues that:

[5] Though not, of course, if the original observations of the phenomena prove to be erroneous.

> “[I]n order to decide whether a particular empirical result supports or confirms or corroborates a particular theory the way in which that theory was developed or constructed needs to be known – more especially, it has to be checked whether or not that empirical result was itself involved in the construction of the theory” (Worrall, 1985, p. 301).

The same basic idea is also, as pointed out above, invoked in Leplin’s Independence Condition. Recently, following Nickles (1987) and Mayo (1991), this view has been termed the “use-novelty” (UN) account, and I follow this terminology here. What each version of the UN account has in common is that it is irrelevant whether an empirical result predicted by a theory is observed before or after the theory is proposed, and that it also does not matter what other theories were being considered at the time of the observation. The result confirms the theory just in case it is predicted by the theory but the theory was not designed to have that result as a deductive consequence.

Worrall (2000; 2002; 2006; 2010a; Scerri & Worrall, 2001) has given a more precise characterisation of the UN account *via* the notion of parameter-fixing. He argues that theory confirmation proceeds in two stages. The first stage can be thought of as a form of “deduction from the phenomena”. Consider a theory with a single free parameter. A parameter here is not necessarily to be understood as a numerical parameter, but more generally as a factor that may take various potential values and thereby partially determine the observational consequences of the theory. This theory’s relationship to some class of empirical results is such that, assuming it is true, the parameter can only have one possible value. As an example, consider Newton’s law of gravity, which states that the gravitational force (F) between two objects is given by:

$$F = \frac{Gm_1m_2}{r^2}$$

Where $m_1$ and $m_2$ are the objects’ respective masses, r is the distance between them, and G is a constant. From the perspective of this general law, the constant G is a free parameter. The Michell-Cavendish experiment is the classic means of determining the value of G (Cavendish, 1798; McCormmach, 1998) [6]. Without going into detail, the outcome of this experiment is a measurement for each of the terms in the above equation except G. Any given experimental result therefore corresponds to one and only one possible value of G. A specific version (i.e. with the free parameter now fixed) of the general gravitational theory is thereby *deduced* from the experimental result. It is this result that the specific version of the theory is "constructed" to accommodate.

The deduction of a specific theory cannot provide confirmation for this general law, since it presupposes its truth. This, however, is the starting point for the second stage. The specific version of the theory deduced from the phenomena may entail some additional empirical consequences that turn out to be correct. For instance, once a value of G is determined by the Michell-Cavendish experiment, the specified theory correctly entails the precise quantitative results of various gravitational interactions. These additional empirical results count as UN predictive successes for the general theory, and so confirm it. Notice that it is sufficient but not necessary for use-novelty that these additional results be temporally novel with respect to the parameter-fixing step. The temporal novelty account therefore emerges as a special case of the UN account.

There are several major objections to the UN account. One, by Harker, is relatively easily dealt with. The others are discussed in the following three sections. Harker favours a version of the theoretical novelty view. He claims that, in Worrall's defence of the UN account, Worrall "appeal[s] to phenomena that [were] either anomalous or at least unexplained by rival theories" (Harker, 2008, p. 445). So all cases where a theory was predictively success under the UN criterion are *simultaneously* cases where the theory was successful under the theoretical novelty criterion. Harker then argues that it is the striking success of these theories

[6] The version of gravitational theory Cavendish used did not actually include such a constant, and he published a figure for the density of the earth. The experiment however *can be*, and has been, used for this purpose.

over their rivals that really motivates the increased degree of confirmation we attribute to these theories.

Harker's objection is doubly mistaken. The first mistake was discussed in the previous section, in response to the theoretical novelty accounts of Musgrave and Leplin. Like these accounts, Harker's objection fails to distinguish between a scientist in some particular epistemic situation having a reason to *prefer* a theory to a rival versus scientists in general having a reason to regard it as *true*. So Harker's objection fails to address the main appeal of the UN account, namely its ability to pick out those cases where the NMA is potentially applicable. This leads on neatly to the second mistake in Harker's objection. While it is certainly true that that Fresnel's theory, for instance, is counted as predictively successful under both the UN and the theoretical novelty accounts, it is incorrect to state that the two accounts *always* agree about which theories are successful. There are several historical cases where the two diverge, and where the UN account more accurately captures the judgements of working scientists.

First, consider the well-known example of the Copernican Revolution of the sixteenth century (see Chapter 4.4 for a detailed exposition of this case study). Copernicus' heliocentric (or heliostatic) model famously (Kuhn, 1957) did not entail any apparent planetary motions which the Ptolemaic geocentric (or geostatic) system did not. Both accounted for the astronomical data available at the time with roughly the same degree of accuracy. So the Copernican model does not enjoy any theoretical novelty over the Ptolemaic one. How, then, to account for the fact that some scientists (notably Kepler and Galileo) adopted the new system immediately? For Worrall (1990), the answer is that several qualitative features of the data set, notably the stations and retrogressions of the planets, emerge naturally from the basic Copernican postulates, but have to be added in "by hand" to the Ptolemaic system. These phenomena, in other words, are *use-novel* in respect of the Copernican, but not the Ptolemaic, system. Thus, if Worrall's reasoning is sound, this is at least one case where the theoretical novelty and UN accounts do not coincide and, in this case, the best scientists (immediately) and the general scientific community (eventually) favoured the theory that was superior under the UN account.

The Copernican Revolution represents a case where neither theory was theoretically novel, but one was use-novel. The divergence between the two accounts is further demonstrated by cases where none of the theories in competition was use-novel, but one was theoretically novel. One set of cases that satisfy this criterion are those where a theory has been modified in an *ad hoc* way to accommodate some recalcitrant empirical phenomenon. Since the newer version of the theory, by definition, accounts for an empirical phenomenon that the earlier version does not, it achieves theoretical novelty. Yet it is not at all clear that we should regard this newer theory as better-confirmed than the older one. This sort of example is, however, arguably somewhat unfair to Harker, as he explicitly advocates simplicity as an additional theoretical virtue, and the sort of ad hoc modification suggested in this example is very likely in violation of this virtue. Nevertheless, there is a difference in that Harker must cite simplicity as an *additional* desideratum, whereas the disapprobation of *ad hoc* manoeuvres is an integral part of the UN account (see sections 8 and 9 for an extensive discussion of empirical virtues of simplicity and unity). To conclude, therefore, Harker has not been able to show the superiority of the theoretical novelty over the UN account. Quite the contrary, where the two diverge, the UN account seems to account for the judgements of actual scientists more naturally.

## 6. Severe testing

Several authors, most notably Mayo (1991; 1996; 2008) and Howson (1984; 1990) have argued that the UN criterion of theory-confirmation is ultimately subordinate to a more fundamental criterion, that of having passed a severe test. Mayo takes inspiration from Popper’s idea that a theory is maximally corroborated by testing it in cases where it is *prima facie* likely to be refuted.

> “A theory is tested not merely by applying it, or by trying it out, but by applying it to very special cases – cases for which it yields results different from those we should have expected without that theory, or in the light of other theories. In other words we try to select for our tests those crucial cases in which we should expect the theory to fail if it is not true.” (Popper, 1963/2002a, p. 150)

Mayo extends this idea, arguing that a hypothesis (h) should be regarded as well-confirmed just in case it passes a severe test (T), where a severe test is defined as one in which “[t]here is a very high probability that test T would not yield ... a passing result, if h is false” (Mayo, 1991, p. 529).

This requirement of severity can be cashed out more precisely with reference to Bayesian confirmation theory. As Howson puts it:

> “But the most important consequence of the Bayesian theory is the fundamental principle of all inductive inference: evidence supports a hypothesis h the more, the less it is explicable by a plausible alternative compared with its explicability by h.” (Howson, 1990, p. 226)

To interpret this statement, recall the Bayesian updating formula already discussed in section 3:

$$p_{new}(H) = p(H|E) = \frac{p(E|H) \times p(H)}{p(E)}$$

In many cases, the degree to which a hypothesis is confirmed by a particular piece of evidence (i.e. the degree to which p(H|E) is greater than p(H)) will depend crucially on the value of p(E). The more improbable the evidence E on the assumption that the hypothesis is false (i.e. that some plausible competing hypothesis is true), the greater the degree of confirmation that results from an observation of E.

These authors go on to argue that there are cases of hypotheses accommodating empirical observations in such a way that they do not satisfy the UN criterion, and that these instances of accommodation nevertheless represent “maximally severe” tests[7]:

> “Passing a Maximally Severe ... Test: h passes a maximally severe test with e iff there is no chance that the test yields such a passing result, if h is false.” (Mayo, 1991, p. 530).

These hypotheses are, in other words, completely confirmed by the severity account, are accepted as confirmed by scientists and our own intuitive judgements, and yet are not at all confirmed under the UN account.

The standard illustrative examples for this point are cases of statistical estimation or simple arithmetical calculation. An arithmetical example is developed by Mayo (1991). Consider a hypothesis, H, with a single free parameter x. It states:

> H(x): The average SAT score of students in a particular class = x

The appropriate value at which to fix this parameter is derived simply by adding up the scores of all the students and dividing by their number, i.e. by calculating the average. Say that the average is some particular value k. The following hypothesis is thus derived:

> H(k): The average SAT score of students in a particular class = k

The students’ actual results are used in deriving this specific hypothesis from the more general version, and so the use of this evidence to test the theory would

[7] Mayo, in fact, also argues that there are cases of theories successfully predicting empirical observations in such a way that they satisfy the UN criterion and utterly failing to satisfy the severity criterion. These cases have not, however, been taken up in the subsequent literature, and so I put them to the side here.

certainly not be use-novel. There is, however, no better test for hypothesis H(k) than simply recording the these results and making the relevant calculation. Because H(k) follows deductively from the evidence, there is no chance that it would pass this test if it were false. The test is therefore, according to Mayo, maximally severe.

A statistical example is developed by Howson (1990). Consider, he suggests, an urn which contains white tickets and black tickets in some unknown proportion. Some number of tickets is drawn at random, resulting in a ratio r/k of black tickets to total tickets drawn. Now consider the following hypothesis about the proportion of black tickets in the population, p:

$$p = r/k \pm \varepsilon$$

This defines a confidence interval, where ε is an error term partially dependent on k. This hypothesis, Howson argues, is maximally confirmed by the very same data used to construct it.

So there appear to be examples of maximally severe tests that do not satisfy the UN criterion. These examples, however, are easily dismissed as relatively trivial and/or unrepresentative of actual cases of scientific reasoning. Worrall, for instance, argues that:

> “In the deterministic case, we *measure* a parameter (or demonstrate that that parameter has a certain value); in the stochastic case we *estimate* a parameter…. [T]here seems to be in this particular case a very good reason why we do not talk of tests: despite Mayo’s claims, a test of a theory surely must have a possible outcome that is inconsistent with the theory— neither the SAT score process nor the confidence-interval technique could possibly refute the ‘theory’ that we end up with” (Worrall, 2006, p. 58, emphasis in original).

Worrall’s (2006; 2010b) more detailed interpretation of the SAT-type example is also instructive. Recall that his initial exposition of the UN account in the previous section postulates two distinct stages of confirmation. The first stage involves assuming some background or general theory is correct, then using empirical results to derive a particular version of this theory. In the second stage, this specific version is tested against empirical results that are not so used, and the general theory is confirmed just in case the theory passes these tests (i.e. it makes use-novel predictions). Applying this framework, Worrall argues that the cases of measurement or estimation highlighted by Mayo *et al.* are instances of stage-one “deduction from the phenomena”. The ‘background theory’ in the SAT case is nothing more than a collection of mathematical truths. In cases of statistical estimation, there are generally some more substantive background assumptions about what type of statistical model is applicable to the case at hand. Estimation therefore involves deducing the value of a certain parameter in the model. But, importantly, these cases of estimation still only count as “severe tests” for the hypothesis that the parameter has this particular value. They do not serve at all to confirm the more general hypothesis that this sort of model is applicable. We normally do not notice this, however, because we regard the applicability of the relevant model as well-established (especially in the deterministic SAT-type case, where the ‘model’ is an analytic truth of mathematics).

In response to this sort of worry, Mayo emphasises that the examples of simple measurement and statistical estimation are intended mainly to “set the mood” (Mayo, 1996, p. 272) for more realistic counterexamples. It is therefore worth examining an actual historical case study discussed repeatedly by Mayo (1991; 1996; 2010), namely the famous test of Einstein’s theory of general theory of relativity (GTR) carried out in 1919 (the original report of this experiment is Dyson, Eddington, & Davidson, 1920). GTR predicts that rays of light will be gravitationally deflected by massive objects. On earth, this effect is most noticeable for stars whose positions are near to that of the sun in the sky, since light passing close to the mass of the sun will experience the gravitational effect relatively strongly. However, the light of these stars is normally not visible on earth due to the glare of the sun. They can only be observed during a solar eclipse. The 1919 experiment involved capturing images on photographic plates of a particular set of stars both during an eclipse and then on another occasion at night, when the position of the sun was nowhere near that of the stars. The degree to which starlight is deflected by the sun could then be determined by comparing the relative positions of the stars on the photographic plates taken at night and during the eclipse.

As reported by Dyson *et al.*, the result of this experiment was a decisive victory for relativity. Of the three sets of plates collected, two were clearly compatible with the amount of deflection predicted by GTR, but not compatible with either the Newtonian theory of gravity (which predicts a smaller degree of deflection) or with the hypothesis that the light is not deflected at all. Of particular interest to Mayo, however, is the result of a third set of plates, taken at Sobral in northern Brazil, which appeared to suggest a degree of deflection *inconsistent* with GTR but consistent with the Newtonian hypothesis. Dyson *et al.*, however, judged that the measurements from this set where untrustworthy, on the grounds that these images were visibly out of focus (describing them as “diffused”). They suggested that this alteration of the focus was probably due to “unequal expansion of the [telescope] mirror due to the sun’s heat” (*ibid.*, p. 309). Mayo argues that this reasoning is manifestly a violation of the UN criterion, since the same piece of evidence, namely a particular set of images, both stands as a potential counterexample to Einstein’s hypothesis and is used to construct a modified version of the hypothesis (i.e. that this particular set of images is not trustworthy) which avoids the counterexample.

Mayo also argues that this case does not appear to match the “deduction from the phenomena” schema.

> “[T]he data-analytic methods, well-known even in 1919, did not assume the underlying theory, GTR, nor is it correct to imagine arguing that, provided you accept GTR, then the mirror distortion due to the Sun’s heat explains why the 1919 Sobral eclipse results were in conflict with GTR’s predicted

> deflection… GTR does not speak about mirror distortions… It was clear the plates, on which the purported GTR anomaly rested, were ruined; accepting GTR had nothing to do with it. Nor could one point to GTR's enjoying more independent support than Newton[‘s theory of gravity] at the time – quite the opposite.” (Mayo, 2010, p. 160)

In other words, it is not the case that the general version of GTR was regarded as well-confirmed, and the anomalous data served merely to derive a specific version of this theory that included mirror distortions. Rather, the *same piece of evidence* is used both to construct and confirm this specific version of the theory.

Mayo's interpretation of this case is seriously flawed. Firstly, it is not at all clear that “GTR does not speak about mirror distortions”. It is a commonplace of philosophy of science, following Duhem, that theories are not tested in isolation. An abstract theory only results in a concrete empirical prediction when conjoined with various auxiliary assumptions, including those related to the reliability of measuring instruments. And many of these may be *tacit assumptions* which are assessed on their own merits only when a prediction fails. It seems clear, therefore, that the *overall* theory or collection of theories being tested in the eclipse experiment does include the tacit assumption that the telescope mirrors are not distorted, and it is precisely this assumption which is challenged by the recalcitrant evidence.

Even so, Mayo may argue, surely it is still the case that specific version of the theory derived by altering the assumption that the focusing mirrors are undistorted is both constructed and confirmed by the same piece of evidence? This argument rests on an elementary confusion. It is certainly true that there is only one “piece of evidence” under discussion – the questionable images. Nevertheless, there are clearly several distinct *facts* observable from this “piece of evidence”. One fact is that the measured relative positions of the stars are contrary to what would be expected under GTR. Another fact is that the image is “diffused”. If Dyson *et al.* had proposed the mirror distortion version of GTR simply on the basis that one set of images gave an answer divergent from the standard version of GTR, their reasoning would, quite rightly, have been rejected. That the image displayed some *independent* evidence of being unreliable – i.e. being visibly out of focus – is crucial. These scientists' background knowledge that this type of mirror tends become distorted by the sun's heat, moreover, provided credibility to the hypothesis that distortion might occur under the circumstances of the experiment. So, applying a more fine-grained notion of a 'unit' of evidence, therefore, Mayo is incorrect to state that the same piece of evidence is used in both constructing and confirming a specific version of the theory.

The specific counterexamples directed by proponents of the severity view against the UN account are thus refuted. There is also, however, a more general criticism of the severity view. Recall that the severity of a test is crucially dependent on the probability of the evidence on the assumption that some "plausible alternative" to the theory in question is true. But it might well be asked what, precisely, allows one to designate one or other alternative theory as "plausible". On the one hand, if plausibility is relativised to a particular historical period, the severity account seems to degenerate into the theoretical novelty account. A well-confirmed theory is one which predicts phenomena that competing theories do not or, better, are unable to explain 'in a natural way' at all. But, under this interpretation, the severity account is now subject to the various criticisms directed at the theoretical novelty account in section 4. Specifically, it seems to depend on an ineliminable element of historical contingency. And historically contingent criteria tell us, at best, which theory is to be preferred in a particular epistemic circumstance. They cannot tell us which theories are likely to be *true*.

If, on the other hand, "plausibility" is to be judged by reference to 'objective' criteria, the severity account is in effect dependent on a more fundamental account of confirmation. After all, if it was readily apparent which theories are 'objectively plausible', we would not be worried about severe testing at all. In the following sections, I will argue that various accounts focused on the theoretical virtue of 'unifying power', including later developments of the UN account, offer such an objective criterion. The notion of a severe test, while possibly still heuristically useful in designing experiments and so on, is ultimately subordinate to these more general accounts of confirmation.

## **7. Weak predictivism**

Musgrave's distinction between logical and historical theories of confirmation was addressed in section 2. In subsequent sections, particularly section 4, it has been argued that historical theories of confirmation are inadequate for the NMA. Historical accounts can, at best, give grounds to prefer one of several competing theories, and so cannot give grounds for regarding any of these theories as likely to be true. In this section, following Hitchcock and Sober (2004), I introduce another distinction, between strong and weak predictivism. Strong predictivism claims that a theory's prediction of some phenomenon is, by itself, reason to regard that theory as more confirmed than another theory which merely accommodates the same phenomenon. Weak predictivism, on the other hand, regards predictive success as merely a "symptom" that some other, more fundamental, epistemic virtue has been satisfied. This virtue could also be fulfilled in certain cases of *post hoc* accommodation. Just as our adequacy condition weighs in favour of logical accounts of empirical success, in this section it will be argued that it also weighs in favour of weak predictivism.

Sober and Hitchcock argue that "*overfitting"* a body of evidence is a cardinal sin of theory construction. And they endorse weak predictivism on the grounds that successful prediction is a sign that a theory has avoided this cardinal sin. To illustrate this idea, Hitchcock and Sober (2004) use the example of a simple curve fitting problem. Suppose that the dots in Figure 2 represent a series of data points generated by some data-generating process. The aim of the exericse is to find a function that accurately describes the empirical data resulting from this process, past and future. For each proposed function, it is standard to characterise "goodness-of-fit" by the sum of the squares of the y-axis distance between each data point and the corresponding value of the function. Finding the best-fit function from some predefined class of functions is therefore often referred to as the “method of least squares”. Notice, however, that it is always possible to find a function that fits any data set *perfectly*, so long as this function is sufficiently complex (i.e. has enough free parameters).

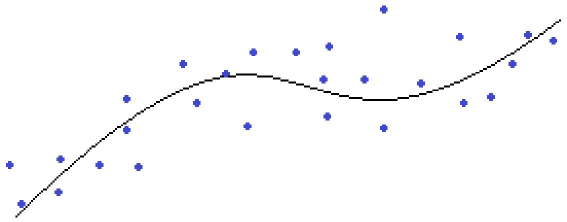

**Figure 2.** A function fitted to a data set

Hitchcock and Sober argue that perfectly accommodating (“overfitting”) a particular data set in this way is almost always ill-advised. This is because most data sets that arise from real observations are “noisy”: the result of each measurement is *largely* determined by the underlying data-generating process, but various idiosyncratic factors such as measurement errors or processes other than the one of interest also contribute to the measured result. If a function perfectly accommodates every data point, then, a lot of what it is being accommodated are these idiosyncratic factors. But subsequent measurements of the same basic process will usually not be influenced by these factors, or at least not in a consistent way. So a function overfitted in this fashion will predict the values of these additional measurements less accurately than a simpler function would.

It is worth emphasising that the assumptions underlying Hitchcock and Sober’s argument are not insubstantial. Given a set of empirical observations, it is not at all trivial to assert that there *is* a unified data-generating process responsible for it. Moreover, it is certainly not a *prima facie* truth that all real data sets are noisy. Nevertheless, these assumptions accurately characterise many cases in empirical science, and have the status of methodological generalisations. If a physicist is presented with a ‘black box’ whose interior is mysterious but which has quantifiable inputs and outputs, these are typically the assumptions she will apply, at least in the first instance, as she attempts to characterise its behaviour.

Avoiding overfitting, in this sort of case, really comes down to selecting the appropriate family of functions (called a “model”) for representing the data set. The linear functions together comprise one example of a model, as they can all be represented by the generalised formula “*y = a + bx*”, where x and y are variables representing data values and *a* and *b* are free parameters. Once the model is selected, the precise form of the function is completely determined by the existing data points, as the specific function is obtained by taking whichever parameter values give the closest fit. In general, model selection begins with a very broad class of functions, from which one particular model is chosen. For instance, many data sets can be fitted by a polynomial of sufficiently high degree, so the model selection step in these cases amounts to deciding which higher-order terms of an arbitrary polynomial to eliminate.

Recall that the aim of this exercise is, given a particular data set, to find a function that will predict the values of not only the existing data points, but also those of *additional* data points resulting from the same underlying data-generating process (if one exists). One intuitive way to achieve this is by ‘directly testing’ the ability of various functions to achieve predictive success. In the context of a curve-fitting problem, this is carried out in a rigorous way by a technique called “cross-validation”. This involves partitioning the given data set into two, at random, fitting various functions against the first partition (the “training set”) and then testing their predictive accuracy against the second partition (the “validation set”). This process can then be repeated for different random partitions of the data set, eventually yielding an aggregate cross-validation score for each type of function.

However, assuming that there is indeed an inverse relationship between predictive power and overfitting, predictive power can also be improved by tweaking model selection specifically to avoid overfitting. Recall that to fit a given data set perfectly requires a model with a large number of free parameters. So, argue Hitchcock and

Sober, a function that is likely to be predictive can be chose by *explicitly* trading off between fit to the existing data and the number of free parameters. Under this sort of approach, each model under consideration is fitted against the *entire* given data set. It is then assigned a score (according to a certain “information criterion”) that increases with goodness of fit of the final fitted function and decreases with the number of free parameters it contains.

There are many additional formal and statistical issues around model selection, but so far as the main thesis is concerned, the main point is that, if Hitchcock and Sober are correct, there is no intrinsic reason to prefer predictively successful models. The primary sin is overfitting, and this can be avoided either by testing a function’s predictive power, or by determining a score for it based on an information criterion. Extending this approach to scientific theories more generally, a theory can be confirmed either by novel predictive success, or by accurately accounting for a large amount of evidence without introducing many adjustable parameters. Hitchcock and Sober therefore advocate a form of weak predictivism. It should also be noted that Hitchcock and Sober’s approach is, under Musgrave’s classification, a logical theory of confirmation. For them, a theory’s degree of confirmation is primarily dependent on the amount of empirical data it accurately explains and its intrinsic simplicity (i.e. paucity of free parameters), neither of these factors being historically contingent.

Interestingly, Hitchcock and Sober explicitly frame their position as a form instrumentalism or empiricism:

> “In the version of the problem that we will address, the goal of scientific theorizing is predictive accuracy. Thus our formulation of the problem is instrumentalist in character. There is a sense in which a theory whose predictions accord well with observation is 'approximately true', and we do want a theory that is approximately true in that sense. But we are not looking for a theory that is approximately true in any deeper sense. In particular, we do not care whether the theory we adopt correctly identifies the degree of the true

> polynomial (for that matter, we do not care whether the true curve is a polynomial at all)." (Hitchcock & Sober, 2004, p. 10)

Notice the parallel here with van Fraassen's claim, discussed in Chapter 1.8, that the aim of science is to obtain theories that are approximately true *only* in respect of observable phenomena. Recall, however, the realist response to this sort of instrumentalism – that the ability of a theoretical abstraction to latch onto an empirical regularity is an achievement that demands an explanation. And the best explanation is simply that the theory is also true of unobservable phenomena (or, in the case discussed above, that the model correctly identifies the "true polynomial"). Hitchcock and Sober's account therefore does not amount to an *argument* either for or against realism. It can be given either empiricist or realist interpretations, depending on whether we are prepared to accept the validity of inference to the best explanation as a pattern of argumentation.

Several other authors have advocated a similar approach to Hitchcock and Sober. Lipton (2004, pp. 168-184) argues that a theory is more strongly supported if it successfully predicts some phenomenon than if it accommodates that phenomenon because with accommodation there is always the risk that the theory has been "fudged". A fudged theory is one in which the felt need to accommodate some known phenomenon or phenomena has induced the theorist to forego theoretical virtues such as simplicity and independent testability. So Lipton in effect also advocates a form of weak predictivism, as he thinks that prediction is a *sign* that a theory is likely to have fulfilled certain more fundamental epistemic requirements.

Maher (1988; 1993) has given an argument which purports to favour UN predictivism but, as argued by Lange (2001), in fact supports a form of weak predictivism. In support of his view, Maher offers a thought experiment, which I present in modified form here. Suppose that a coin will be tossed 100 times and the outcome, head or tails, recorded. An experimental subject will be asked to predict the outcome of the 100th toss. Now imagine two cases. In the first, the subject observes the first 99 tosses, which give an apparently random sequence of heads and tails, then states that the final toss will be heads. This prediction turns out to be correct. In the second case, the subject announces after only 50 tosses a prediction for all the remaining tosses. This prediction, including that for the 100th toss, also turns out to be correct. Assume that the actual sequence of results is the same in each case.

In both cases, the subject of the coin-tossing experiment has proposed a ‘hypothesis’ correctly stating the outcome of the 100 coin tosses. Yet, intuitively, the hypothesis is more confirmed by success in the ambitious prediction than it is when it successfully predicts the result of only the last toss. This meagre prediction, after all, has a 50-50 chance of coming out correct simply by accident. Maher accounts for this intuitive judgement as follows:

> “Clearly the reason for our different attitude in the two situations is that the successful prediction is strong evidence that the subject has a reliable method of predicting coin tosses, while the successful accommodation provides no reason to think that the subject has a reliable method of predicting coin tosses.” (Maher, 1993), p. 330)

So it is not that there is any intrinsic difference between the hypotheses (which is just as well, since they are, by assumption, identical in this thought experiment), but a difference in the *method* by which the hypothesis was derived in each case.

Lange (2001) “tweaks” Maher’s argument by asking us to consider the case where the results of the coin tosses are not random, but rather follow some regular sequence, say alternating tails and heads. Now suppose we again have two cases, one where the subject accommodates most of the data, and one where she predicts a large portion of the data. Lange claims that the intuitive difference in confirmation between these two cases now disappears. What this shows, he argues, is that in Maher’s original case, it is the apparent lack of connection between the different cases subsumed by the hypothesis (i.e. the results of independent, random coin tosses) that leads us to regard it as ill-confirmed. The hypothesis comprises an “arbitrary conjunction”. When the subject successfully predicts the results of a “random” sequence, we infer that she has uncovered some regularity underlying these results, perhaps one we have simply not noticed. When the subject merely accommodates these results, however, we have no reason to doubt that she has simply lumped them together into an arbitrary conjunction. In the case where the pattern is ‘obvious’, in contrast, there is no such difference, as we infer that both the predictor and the accommodator have discovered it.

So, according to Lange, the fundamental epistemic vice is arbitrary conjunction. Accommodation is *sign* that a theory has potentially been constructed in this way because of the method by which scientists typically go about constructing theories:

> “[An arbitrary conjunction] is unlikely to be formulated until after a great many of its cases have already been checked [i.e. accommodated]. That is because a few examined cases do not suffice to lead scientists to formulate that particular hypothesis... This association makes it *appear* that predictions are somehow intrinsically superior to accommodations, when in fact that an observation was predicted rather than accommodated does not *make* that observation better evidence for the hypothesis.” (*ibid.*, p. 577)

Lange’s account, however, surely requires an explanation of what is so bad about arbitrary conjunctions. In providing this explanation, Lange quotes Goodman:

> “Consider the heterogeneous conjunction:
>
> 8497 is a prime number and the other side of the moon is flat and Elizabeth the First was crowned on a Tuesday.
>
> To show that any one of the three component statements is true is to support the conjunction... But support of this kind is not confirmation; for establishment of one component endows the whole statement with no

> credibility that is transmitted to other component statements.” (Goodman, 1983, pp. 68-69)

This is suggestive, but still not definitive. Both Lange and Goodman seem in fact to be tacitly appealing to the theoretical virtue of simplicity or ‘unity’. In a unified theory, there are some basic principles which account for much of the seemingly unrelated empirical data. In theories constructed by arbitrary conjunction, there are no such basic principles. This is why confirmation of one part of the theory does not confirm the theory as a whole; there is no principle by which confirmation might be “transmitted”.

Goodman and Lange are, of course, not alone in the intuition that the ability to unify diverse phenomena is one of the basic, or *the* basic, epistemic virtue that a scientific theory can possess. Whewell, for instance, argues that the “best established theories [are those where] inductions from classes of facts altogether different have ... *jumped together*” (Whewell, 1858/1968, p. 153, emphasis in original). More recently, Friedman (1974) and Kitcher (1981; 1989) have argued that scientific explanation should be understood as essentially involving unification. I shall not here attempt to defend the primacy of unifying power as a theoretical virtue (although I discuss this idea further in section 8 of this chapter and in Chapter 4). For now, simply note that this idea is intuitively not implausible, and that it animates the various forms of weak predictivism surveyed here.

Harker also implicitly appeals to the virtue of unifying power. He criticises the UN account as follows:

> What is significant about verified forecasts, aside from their being explained by only one theory, is not that those data weren’t *used*, but that no additional assumptions were required for the explanations of those data. The successful theory enjoys increased explanatory strength without any loss of theoretical simplicity.” (Harker, 2008, p. 447, emphasis in original)

Notice how similar Harker’s approach is to the sort of weak predictivism advocated by Hitchcock and Sober. Whereas Harker discusses a trade-off between explanatory strength and simplicity, Hitchcock and Sober argue that a theoretical model should achieve some optimal balance between goodness of fit and the number of free parameters it contains. In both cases, the central concern is that the complexity of the theory be appropriately ‘matched’ to the complexity of the data set it is to explain.

In this and the previous section, I have surveyed some of the forms of weak predictivism on offer. The basic claim of weak predictivism, recall, is that prediction is not intrinsically superior to accommodation as grounds for accepting a theory, but that it frequently gives good evidence that some more fundamental epistemic desideratum has been satisfied. But there are various different, albeit interrelated, proposals on what this fundamental epistemic requirement *is*. Mayo and Howson (see section 6) think that a theory is most highly confirmed by a certain sort of *accomplishment*, namely passing a severe test. Making a use-novel prediction sometimes counts as passing a severe test but, they claim, not all severe tests are novel predictions. Hitchcock and Sober claim that the degree to which a theory is confirmed is fundamentally dependent on how it is (or, more precisely, is not) *constructed*; i.e. it must have been fitted to the data, but not overfitted. Lange and Lipton think that the fundamental epistemic virtue is *intrinsic* to the theory or, at least, to some combination of it internal structure and its empirical consequences (think, for instance, of “independent testability”). Harker splits the difference, arguing that the fundamental epistemic virtues consist of both intrinsic features (such as unity) and accomplishments (such as accounting for phenomena that rival theories cannot).

The judgement that a theory has passed a severe test, or that it is more empirically successful than a rival theories, will, as I have argued previously, always depend on what other background or rival theories are present. This historical contingency results means that, while the accomplishment model of confirmation may be able to determine which of a set of rival theories is to be preferred, it will never give us good grounds to believe that a given theory (or part thereof) is *true*. It is thus inadequate the aim that has been set out.

The construction and intrinsic features views are, it seems, two sides of the same coin. To claim that a building, say, has certain features is just the same as claiming that it has been constructed or modified in a certain way. And, having discarded accomplishment versions of weak predictivism, all the remaining versions appear to be animated by the same underlying logic. Lange argues that arbitrary conjunctions are to be avoided. Hitchcock and Sober claim that a theory is likely to be successful just in case it achieves the optimum balance between goodness of fit and number of free parameter. Harker argues that a theory must maintain the optimum balance between explanatory strength and simplicity. Although they express it differently, all these different models of weak predictivism regard *unifying power* as the fundamental epistemic virtue.

A form of weak predictivism that emphasises unifying power may seem to represent a challenge to theories of novel prediction generally, and the UN account specifically. In the following section, however, I shall argue that a very reasonable development of the UN account due to Worrall is in fact committed to a kind of unification view. Using Worrall's version of the UN account, I shall also show how the theoretical virtue of unifying power may be applied in a theory of confirmation and how such a theory may be adequate to the NMA.

## **8. The unification view**

A common objection to the UN account is the so-called "twin paradox", variations of which are discussed by Hitchcock and Sober (2004), Lipton (2004, ch. 10) and Musgrave (1974). Returning to the example discussed in section 5, suppose that a pair of twins, A and B, are both independently working on Newton's theory of gravity. Twin B carries out a Cavendish-type experiment, $E_1$, and uses the measured result to fix G at a particular value, say $k_1$. The specific theory thus derived now accounts not only for the result of $E_1$ but also correctly predicts the result of another experiment, $E_2$. Twin A takes measurements for both $E_1$ and $E_2$ and uses *both* results to fix G = $k_1$. A does not predict the result of any additional experiments. Since B makes a successful UN prediction and A makes none, it seems we must conclude that the theory is better confirmed for B than it is for A, even though each has adopted the same theory and has the same evidence for this theory. Not only is the conclusion absurd on its face, it also renders indeterminate to what extent the realist should now view the theory as likely to be true. The reason this paradox is possible is that focussing on which of the empirical results a given scientist *uses* introduces an irreducible element of historical contingency. So the UN account, at least as initially formulated, is also susceptible to the objection which had led us to reject the temporal and theoretical novelty accounts.

Worrall denies that this imagined situation is paradoxical, arguing instead that "scientist B has *shown*, what A simply subjectively failed to recognise, that the theory is supported" by the measurement not used in fixing the parameter (Worrall, 2006, p. 54, emphasis in original). The twin paradox has bite only on a common-sense reading of the term "use" as implying *intent*. For instance, I *use* a hammer only if I intend to hit a nail, but not if I accidentally drop it onto the nail. By analogy, the suggestion is that a scientist *uses* a result in designing the theory if and only if she intends that the theory be able to account for it. And it seems plausible that a scientist like A could intend to account for both results when setting the parameter, and so use them both. However, according to Worrall, this is a mistaken reading of "use" under his version of the UN account. It is simply a logical feature of the general theory that one and only one experimental measurement is *required* to fix the single free parameter G to obtain the appropriate specific theory. The theory is thus equally well confirmed in either circumstance.

Worrall's goes on to consider the situation where, as before, the specific theory correctly entails observational consequences $E_1$ and $E_2$. However, twin A now derives the specific theory from the general theory *using* $E_1$ and the specific theory then predicts $E_2$; whereas twin B *uses* $E_2$ and predicts $E_1$. Worrall's response is as follows:

> "These may be strictly different accounts but they are surely equivalent modulo any genuine interest that we would have in making confirmation judgments: each of A and B has shown that the general theory needs to fill in one parameter value on the basis of one piece of data, thus producing a specific theory that gains genuine empirical success from the other piece of data (at least—there may of course be other results that the specific theory also correctly predicts). So each scientist shows that there is, so to speak, one unit of genuine, unconditional, general-theory-involving data..." (ibid, p. 55)

So it makes no sense to talk about a circumstance in which one observational result is used and the other predicted, and another in which the opposite holds. Rather a theory is confirmed to the extent that it entails verified empirical content greater than that needed to fix all its free parameters.

In responding to the twin paradox and similar cases, Worrall has moved decisively away from defining "novelty" in terms of what is *used*. Instead, he defines it in terms of what is logically required, or what *needs to be used*. Although Worrall still uses the term "use-novelty" to describe his view, it is therefore a very different notion of novel prediction to that employed by Zahar and Leplin, and much closer to the types of weak predictivism – advocated, as discussed above, by Hitchcock, Sober, Harker, Lange, Lipton, and others – which emphasise the unifying power of a theory. The idea of ensuring that the verified empirical content of the theory exceeds that needed to fix free parameters is very similar to Harker's later idea that there is an optimal balance between explanatory strength and simplicity, for instance. So, to avoid confusion, I shall not refer to Worrall's account as a form of the UN account, but describe it as a version of the "unification view" (UV). For clarity, the particular version of the UV that follows from Worrall's UN account can be restated, as follows:

> A theory is confirmed just in case, and to the extent that, it entails verified empirical content greater than that needed to fix all its free parameters. (UV1)

As further complications are discussed in the following two sections of this chapter, this definition will be amended somewhat. However, for current purposes it gives a good general idea of what is being proposed.

It is worth emphasising again how the UV differs from the simple UN account that appears to be expressed by Zahar (1973). The latter relativizes the degree to which an empirical result confirms a theory to the *actual historical fact* of whether that result was used in constructing the theory. It is, to use Ladyman's (1999) words, concerned with the "provenance of the theory". The UV, in contrast, is not concerned with which *particular* empirical results are used in constructing the theory, only with whether the theory entails more results than are required to construct it. Thus, although it is not obvious in his earlier writing, at least the later Worrall endorses a form weak predictivism. The UV does not give any special status to novel prediction as such, under any natural interpretation of the word "novel".

The UV, then, is a non-historical account of confirmation. Does it satisfy Leplin's relation requirement, that confirmation be a two-place relation between theory and empirical evidence? Worrall claims that:

> "... the 'heuristic' view I advocate ... make[s] confirmation a three-, rather than two-place relation. But, although describable in a loose way as making confirmation dependent on background knowledge, in fact this account makes confirmation (or rather both kinds of confirmation) depend on evidence e, specific theory T´, and the underlying general theory T." (Worrall, 2006, pp. 55-56)

Since theories T and T´ are both characterised in abstract logical terms (say as sets of propositions), there is nothing in this account of confirmation that is historically contingent or indexed to a particular time. So UV satisfies the motivation behind the relation requirement, if not the exact version of it articulated by Leplin. Moreover, notice that given a general theory T and the set of evidence by which it is considered confirmed under the UV, it is always possible to derive the specific theory T´. So Worrall's three-place relation can be reduced, without loss of information, to a two-place relation between general theory T and the body of evidence.

Under the UV, the degree of confirmation of a theory depends only on the intrinsic logical structure of the theory itself and the verified empirical consequences it entails. This means that UV is easily able to account for confirmation due to "old evidence". For instance, the anomalous precession of Mercury's perihelion can provide support to general relativity, since it is part of a larger pool of empirical results which the theory accurately entails. Notice also that UV confirmation is in no way diminished if the theory in question is abandoned in favour of superior alternatives, or indeed if other predictions of the theory turn out to be radically false. As Worrall puts it: "Fresnel's wave theory of diffraction was, is, and forever shall be, confirmed ... by the 'white spot' result—this result follows from that wave theory of diffraction and gives support to the whole wave programme" (ibid, p 56). Indeed, by the same logic, a long-obsolete theory can be confirmed even by entirely new observations made in the present. So the UV is not only able to deal with old evidence, but also provides a particular perspective on 'new evidence'.

This retroactive confirmation of obsolete theories may seem like a counterintuitive consequence of the UV. It is precisely what is desired, however, by the scientific realist. She *wants* to be a realist about at least certain parts of Fresnel's wave theory, since the theory is undoubtedly a successful one. Of course, it would be preferable for this account of novel prediction, *via* the NMA, to confirm only *certain parts* of the theory (and for these parts to be retained in successor theories). How we ought to distinguish "essential" from "inessential" parts of a theory is a question that will be tackled in Chapter 3. For current purposes, it is sufficient to notice that, relative to our current purposes of providing an account of confirmation adequate to the NMA, this consequence of the UV is a definite point in its favour.

Moreover, the intuition that unification provides powerful confirmation for theories is just the same intuition that underlies the NMA (this argument is also made briefly by Worrall (2011). The NMA is motivated by the idea that certain paradigm cases of prediction demand a special explanation, and then seeks to provide an explanation in terms of the truth of the scientific theories involved. But notice that we don't in general demand a special explanation for just any case of "prediction". It is entirely unremarkable if someone accurately forecasts when the train will arrive if they catch the same one every day, for example. What is really remarkable about the paradigm cases is that those making the predictions have so little to go on; they certainly cannot 'read off' the correct result from some robust base of experience. What is 'miraculous' about such cases, in other words, is how much verified empirical content these theories generate compared to the amount that has to be 'fed in' (in the form of results used to fix free parameters). The NMA, in other words, is motivated by the intuition that there is something epistemically special about successful unification. Again, I should be clear that this is not necessarily to state categorically that the NMA is a valid form of argument, but only to argue that, *if* we accept the NMA, then the UV is the appropriate criterion of theoretical success.

It has been argued that the UV is most suited to the realist's aim of determining which theories (or parts thereof) we should believe to be true. This view defines empirical success in terms of the 'excess' verified empirical content entailed by the theory when all its free parameters have been fixed at the appropriate values. In the remaining substantive sections of this chapter, several assumptions implicit in this definition will be challenged, and a somewhat modified version of the view proposed as a result. The first assumption, addressed in section 9, is that constructing a theory is simply a matter of fixing the values of free parameters in a given general theory. This ignores the epistemically relevant question of how this general theory was derived in the first place. The second assumption, addressed in section 10, is that, to be empirically successful, a theory must *deductively entail* empirical results. The final version of UV proposed in section 10 incorporates a broader conception of how a theory may “give rise to” verified empirical results.

## 9. Theoretical unity and theoretical simplicity

According the UV, a theory is confirmed to the extent that it displays “unifying power”. Intuitively, a unifying theory is one which captures many empirical regularities and yet is itself “simple”. This notion of simplicity has, in the previous sections been cashed out more rigorously *via* the idea of free parameters, with the consequence that a theory with relatively many free parameters is referred to as “complex” and one with relatively few as “simple”. Under this schema, a theory is unifying, and so should be considered confirmed, to the extent that it captures more empirical regularities than are required to fix its free parameters. In this section, it will be argued that cashing out the idea of simplicity, and thus of unifying power, by reference to free parameters is not generally correct. Or, at least, it is not correct if the number of free parameters is understood as something that can be determined by simple inspection of a theory. Instead, I will advocate a second version of the UV, which states that:

> A theory is confirmed just in case, and to the extent that, it entails more verified empirical content than that required to *construct* it (where the notion of “constructing” a theory is more general than that of fixing the free parameters that are evident from a simple inspection of the theory). (UV2)

Sober has, with the following argument, very effectively problematised the purported inverse relation between the simplicity of a theory and the number of free parameters it contains. He asks us to consider two models, one linear and one parabolic:

(LIN) $y = a + bx$

(PAR) $y = a + bx + cx^2$

and then introduces a puzzle:

> "Once values for these parameters are specified, a unique straight line and a unique parabola are obtained. Notice that (LIN) is simpler than (PAR) if simplicity is calculated by counting adjustable parameters. However, if simplicity involves paucity of assumptions, the opposite conclusion follows. (LIN) is equivalent to the conjunction of (PAR) and the further assumption that c=0. Since (LIN) is a special case of (PAR), (LIN) says more, not less" (Sober, 2001, p. 17) [8].

Recall the earlier argument that what is "miraculous" about the paradigm cases of novel prediction is that the theory concerned uses only a relatively small number of empirical observations to yield a large number of predictions. What this puzzle highlights is that the face-value complexity of a theory, as ascertained by the number of free parameters it contains, can be misleading about the amount of verified empirical information that needs to be 'fed in' to generate predictions. The background assumptions of the theory embody significant empirical knowledge that has been used at an earlier stage of theorising.

This idea is illustrated by returning to the example of curve-fitting. Taking a somewhat more general perspective than that taken in section 7, a curve fitting problem can be considered in three stages (following Forster, 1988b, p. 67). First, one must select which empirical variables are to be taken as systematically related. In the relatively simple case where only two variables are considered, the putative relationship between them is typically represented as a collection of paired values on a two-dimensional scatter plot. In the second stage, as described in section 7, one selects a model that will be used to represent the general form of this relationship. In the third stage, the model formula is "fitted" to the data set to arrive

[8] A version of this argument is also found in Popper (1959/2002b, appendix *viii).

at a specific function that ‘best’ represents the relationship between the variables. Note that the distinction between these three steps is primarily *conceptual*, and there may be considerable overlap between them in practice. For instance, model selection often involves an algorithm which fits each of the models under consideration to the data set.

Before a scientist begins a curve fitting procedure, she has a data set, but has not yet used any information from it. She thus has no warrant to believe any particular hypothesis about which variables are related to which. Each step in the curve fitting procedure uses information from the data set to obtain a more specific claim about the underlying data-generating process. But there is always some flexibility in whether information should be used in an earlier or a later step. Consider, for instance, the case where it is assumed that the data can be represented by a polynomial. As in Sober’s example, higher-order polynomial terms can be eliminated explicitly during the model-selection step. Alternatively, a higher-order polynomial can be selected as the model, with the higher-order terms then ‘eliminated’ when this model is fitted to the data because the fitting algorithm sets their coefficients to a negligible value. However one goes about the procedure, there is in general a trade-off between using empirical data at an earlier or at a later stage. Or, to put it in Sober’s terms, between making an assumption or leaving a free parameter. It is worth making the obvious point, however, that while empirical information can be used either to inform the assumptions of the model or to fix free parameters, there is no free lunch. The basic architecture of the problem will determine the minimum amount of verified empirical content required to obtain a function that makes successful predictions.

To solve Sober’s puzzle, and the more general concern about theoretical simplicity, a broader picture of the amount of confirmed empirical content used in constructing a theory is required. According to this broader picture, a simpler (and potentially more unifying) theory is just one in which less empirical content has been used in constructing it. This empirical content can be used to fix free parameters, or it can be used to craft theoretical assumptions. Theoretical assumptions, however, can in general be understood as parameters that were fixed at an earlier stage of the construction process. For instance, that a linear function is the appropriate model for fitting a particular data set is an assumption from the perspective of a fitting algorithm, but represents one potential value of a free parameter in the model selection procedure. When one ‘counts’ how many empirical results are required to construct a theory, one therefore needs to count not only the number of free parameters it contains, but also how many previously fixed parameters are embodied as assumptions in the structure of the theory. By direct analogy to the curve fitting example, assessing the complexity of a given scientific theory requires us to understand how the theory is derivable from some more general class of theories by “deduction from the phenomena”, given a set of accepted empirical results.

Curve fitting is an extremely well-defined problem. The broader class of functions from which a preferred curve is to be selected is well characterised, as are the steps by which one might progressively narrow down the reasonable options until a single curve is obtained. Moreover, tools such as information criteria allow one to assess explicitly the degree of unifying power possessed by a fitted mathematical function relative to some data set. For a more qualitative theory, however, the broader class of theories is not nearly so well characterised. And, while one can make intuitive judgements about which theories are more or less unifying, there are no formal tools even remotely analogous to information criteria. So, having defined a more general conception of theoretical unifying power, I shall in the following paragraphs outline some more rigorous principles by which it might be applied to more qualitative theories.

Intuitively, a unifying theory is one for which “you get out more than you put in”. That is, it entails more verified empirical information than is used in constructing it. The amount of empirical information entailed by a theory is, however, generally more obvious than the amount used in constructing it. So the problem at hand is largely one of assessing the latter quantity. As a starting point, consider the difference in theoretical complexity between a general theory and a more specific theory that is derived from it. In going from the former to the latter, some quantum of empirical information must be used to set a free parameter. So it is reasonable to conclude that the latter is more complex than the former by an amount that is directly related to the *minimum* of confirmed empirical content that is required for this derivation. And the specific theory is more unifying than the general just in case this amount is smaller than the amount of additional confirmed empirical content that is entailed.

As an example, consider Fresnel's later wave theory of light. Fresnel's earlier theory claims that light obeys certain general wave equations. Although Fresnel assumed that these waves were longitudinal, the general equations are in fact silent on the question of whether they are longitudinal or transverse (i.e. whether the oscillation of the medium is oriented to the axis of propagation of the wave, or perpendicular to it). However, a question on which a theory is silent is also a tacit free parameter. Given the results of interference experiments with beams of polarised light (conducted with Arago), Fresnel modified his theory to claim that light is comprised of transverse waves (see Worrall, 1990 for a discussion of this episode). Because additional empirical information has been used in constructing it, the specific later theory is more complex than the general earlier theory. It is also more unifying, however, as the expense in theoretical complexity is small relative to the increase in verified empirical content.

So it is relatively straightforward to compare the complexity, and thus the unifying power, of two theories which are related by a single application of deduction from the phenomena. There remain two significant challenges, however. Firstly, nothing that has been proposed so far gives any idea of we might compare theories which are *not* so closely related. And this, of course, is main problem facing actual scientists in the context of theory choice. The second challenge is that, as argued above, the NMA demands an absolute or objective account of empirical success. Even supposing that it is possible to compare very different theories, it does not necessarily follow that a theory which is superior under this comparison can legitimately be regarded as (approximately and/or partially) true. Notice, however, that both problems are solved if the second is solved – being able to place all theories on an absolute scale of empirical success means, *a fortiori*, that any two can be compared. This, therefore, is the problem to which we now turn.

The notion of an absolute account of empirical success is quite plausible in the curve fitting case. This is because this process can, at least potentially, get started without any information about what the final fitted function would look like. The general ‘theory’ in effect is that *some* model will fit the data. This ‘theory’ is so general that it requires no empirical information at all for its construction, thus establishing a ‘zero point’ for the subsequent construction. Given a clearly-defined data set, the process then proceeds, *via* a precise set of algorithms, to a final fitted function. If it is assumed that each step in the algorithm results in an increase in unifying power, starting from zero means that the end point is guaranteed to be empirically successful by some absolute standard. By analogy, determining the empirical success of some theory on the UV would seem to require showing how it is derivable by a series of well-defined steps, each of which increases unifying power, from a theory so general it effectively says nothing at all[9].

Before continuing with this line of thought, however, it is worth considering a few objections and alternatives. One possible alternative is that an absolute standard of unifying power might be defined by reference to particular cases assumed to be paradigmatically successful. For instance, one might say that any theory which is roughly comparable to (or more empirically successful than) Fresnel’s counts as unifying; and any roughly comparable to Ptolemy’s is *not* unifying. A major difficulty with this approach is that comparing paradigm theories to those outside their immediate theoretical ‘neighbourhood’ immediately suggests the proverbial comparison of apples and oranges. The transverse and general wave theories of light are comparable just because it is clear how to go from one to the other in the abstract hierarchy of theories of which they are both members. The same is not

[9] Notice I assume throughout that a theory which does not use any empirical information also cannot generate any significant empirical success. This assumption is based on the more basic assumption that it is impossible (or highly unlikely) simply to ‘guess’ the form of a predictively successful theory without empirical guidance. In many cases of successful theorising, this empirical guidance may of course be tacit or informal, but it is still present.

true with respect to, say, the wave theory of light and Darwin's theory of evolution. A further difficulty is that, even insofar as it is possible compare the empirical success of theories to paradigm cases, the anti-realist surely will not grant that these paradigm cases are (even approximately and/or partially) true. So for the whole enterprise to get off the ground, one must in any case provide additional arguments to show that the paradigm cases are sufficiently unifying by some *absolute standard* that the inference to their truth is warranted.

Another alternative is suggested by Worrall. Responding to an idea related to that described above, he writes:

> "The whole idea of reconstructing our knowledge from bottom up ... is surely a chimera. Surely these justificatory trees grow back into the mists of time to the emergence of homo sapiens and beyond." (Worrall, 2010a, p. 748)

And goes on to state that:

> "It is ... stunning predictive successes that give at least a large part of the credence to the premises from which demonstrative inductions begin—not some stepwise demonstration drifting back into the mists of time." (*ibid.*, p. 749)

So (Worrall argues), we do not, and can never have, sufficiently detailed access to the hierarchy of ever-more-general theories to show how any given particular theory is derived. The project of providing an absolute measure of empirical success is therefore doomed. Moreover, it is not even necessary, because the achievement of (UN) predictive success, plus the NMA, is sufficient for regarding a theory as confirmed.

Although similar positions have been discussed above, it is worth emphasising that Worrall's positive account of theory confirmation is no longer tenable. All the views of empirical success surveyed in this chapter, except for the UV, understand empirical success as an ‘event’ or ‘achievement’. Under the simple UN account, for instance, a theory is confirmed by the event of observing a phenomenon that has not been used in the construction of that theory, but is entailed by it. Under the UV, in contrast, empirical success is a ‘static’ relation that exists between a theory and a body of verified empirical content. Several reasons for rejecting the event model have already been articulated in this chapter. Another reason, which arises from the picture of theoretical complexity outlined above, is that it cannot distinguish between a theory that is empirically successful and one that has simply becomes *less unsuccessful*. To give a simple example, imagine a ‘model’ of a two-dimensional data set which consists of a series of vectors specifying the difference in coordinates between each data point and that following it. This ‘theory’ is manifestly empirically unsuccessful, since for a data set with n points, it has n-1 assumptions. Moreover, it does not correctly entail the coordinates of even a single data point, since the vectors by themselves do not give any information about any actual coordinates. Feed into the model values for a single data point, however, and it will immediately spit out all the remaining values! Under the UN account, this is an amazing success for the theory – only a single data point was used to accurately predict the values for the entire set. Under the UV account, it is clear that, while the addition of a single data point clearly *improved* the empirical success of the theory, it has only improved it to zero – the total set of empirical results used in constructing the theory is exactly the same as the set entailed by theory. The theory as a whole is thus not empirically successful by any sensible (that is to say, absolute) measure.

What this example demonstrates is that a general theory cannot be treated as a ‘black box’. To assess its empirical success against a body of evidence, it is necessary to know not only whether it was empirically successful ‘at the margin’, but how much empirical content was used in constructing it. But the same argument will then go for the general theory from which *it* was constructed, and so on. Otherwise, it could be the case, unbeknownst to us, that any theory in this chain in fact embodies far more assumptions than is commonly supposed, and so renders empirically unsuccessful any theory that is derived from it.

Of course, it might be argued that this is a mere ‘philosopher’s concern’. It is *obvious* that current and past theories considered to be successful are not gerrymandered monstrosities like the model described above. The assumptions of these theories are typically written down clearly and explicitly, and from this we can plainly see that they are usually few in number and general in scope. I share this intuition. But it cannot simply be assumed. Although those assumptions are usually *lexically* simple, they are often the consequence of much hard-won empirical knowledge. Moreover, a theory may embody large amounts of *tacit* empirical knowledge which is not explicitly stated in its assumptions. These tacit assumptions may include, for instance, basic systems of classification or assumptions about causality.

Worrall’s negative argument against absolute views of empirical success is that the required reconstruction of our theoretical knowledge is simply untenable. In response to this, it is worth reiterating that, since the UV demands a ‘timeless’ logical relation between theory and phenomena, what is required is a ‘rational reconstruction’ of the hierarchy. The historical details, where they are available, will obviously assist in this reconstruction, but are not necessary for it.

To illustrate this point, here follows a brief rational construction of the derivation of Fresnel’s theory. It has already been shown that both the longitudinal and transverse wave theories are instances of the nonspecific wave theory. The wave theory in turn is an instance of a more general theory, which also includes particle theories as instances. This general theory claims that light is an entity emitted (or possibly reflected) from the objects of visual perception and is propagated through space at a finite speed before being received by our eyes or other measuring instruments (this is thus generally termed the “reception theory”). As discussed in section 3, it is the results of inference experiments that direct us towards accepting the particular instance of this theory in which the entity propagated through space is a wave. The reception view of light can be contrasted with the view, popular in ancient times up until the early modern period, that the carrier of visual perception is emitted in the first instance *from the eyes* and gives rise to visual experience when it strikes objects (Ronchi, 1957, pp. 24-29). The ‘crucial experiments’ that settle the issue in favour of the reception view are the first successful measurements of a finite speed of light, beginning in 1676 (Romer & Cohen, 1940). This is because the view that the carrier of visual perception emerges from the eyes requires that the propagation of this carrier be instantaneous, to accommodate the known phenomenon that distant objects are perceived the instant we look towards them. Both the reception and emission theories are in turn more specific instances of Euclid’s theory that, all else equal, light consists of “rays” which travel in straight lines, except where they suffer refraction or reflection at the interface between different materials. Because it travels in straight lines, under this theory light is susceptible to geometrical treatment (Clegg, 2008, pp. 20-21). One unifying hypothesis found in simple ray optics is Snell’s law, which systematically relates the angle of an incident ray of light to that of a refracted ray across a wide variety of circumstances.

This reconstruction is intended to provide a sketch of what would be required to assess the predictive success of Fresnel’s theory. It is admittedly incomplete in two important respects. Firstly, it refers only to theories that have actually been considered at various points in history. These historical theories, however, are intended merely as representatives of a more inclusive *logical* hierarchy. Consider, for instance the view that a beam of light originates at a point midway between the observer and observed object and travels in both directions to connect the two. Although this is undoubtedly a strange theory, it is a perfectly sound instantiation of ray optics alongside the reception theory. The broader point is that a free parameter in a general theory may have many possible values, and some of these may correspond to specific theories that have not actually been advocated by anyone. A complete reconstruction would consider the full range of possible parameter values, and show how empirical information can be used to eliminate all but one. Certain theories, incidentally, are rejected out of hand as “strange” because they conflict with empirical results that are too basic to be mentioned in scientists' explicit arguments. Empirical information is nevertheless involved in excluding certain logically possible values of the relevant parameters.

The reconstruction is also incomplete in that it does not terminate at a totally content-free theory. This is largely because it is limited to views that we are inclined to think of as "scientific theories". But as the reconstruction proceeds up the hierarchy of abstraction, it will eventually reach views that are so obvious to us that they do not seem to merit the term. For instance, geometric optics is an instance of the more general view that all the diverse phenomena we group under the term "light" *are*, in fact, instances of a single phenomenon that can be described by a fixed set of laws. This sort of statement is analogous to the statement of which variables are systematically related to each other in the curve fitting case. It seems *prima facie* very plausible that the theoretical hierarchy in each case will terminate with this sort of very basic classification schema. Anything more general represents a refusal to classify the world into types at all, and thus to make any scientific inference. Notice, moreover, that even such a simple system of classification is undoubtedly successful under the UV. We practice science precise because we doubt that the world is merely a blooming, buzzing confusion, and believe that there is an underlying unity to the phenomena. And we accept theories precisely to the extent that they are able to capture this unity.

Before concluding this section, it is worth remarking briefly on the logical relationship between general and specific theories in a particular hierarchy. Since the specific theory is derived by fixing the value of a parameter that is free in the general theory, we can think of the former as resulting from conjoining an additional premise to the latter. The specific theory thus logically implies the general. However, in just the same way that the addition of premises to a set does not invalidate the deduction of any results from that set, each particular theory inherits all the empirical consequences of the more general theory of which it is an instance. This is a striking and possibly counterintuitive result, since many of the empirical consequences of earlier and consequently more general[10] theories have subsequently been refuted. As discussed in Chapter 1, avoid the pessimistic induction requires an account of which elements of theories are "essential" to their empirical success. The realist, therefore, will wish to add this some account of essentialness to the picture presented in this section. The notion of essentialness will be discussed in detail in Chapter 3.

## **10. Deductive entailment and confirmation**

In the previous sections, a unifying theory has been described as one that *deductively entails* some sufficiently broad set of verified empirical consequences. This focus on entailment is a convenient philosopher's shorthand, and gives a good general idea of what is being proposed. However, for a theory to count as unifying, it is in fact neither sufficient nor necessary that it deductively entail empirical results.

Firstly, deductive entailment is not sufficient. A good illustrative example of this point is provided by (Vickers, forthcoming) in his discussion of Velikovsky's (1950) *Worlds in Collision*. In this book, Velikovsky proposes an extremely convoluted history of the solar system which purports to explain, among other things, the supposed occurrence of events described in the Bible. Virtually no serious scientists, at the time of initial publication or since, have accorded any credibility to Velikovsky's theory. Nevertheless, this theory did produce some predictions that were temporally novel and did in fact turn out to be accurate. For instance, on Velikovsky's account, the planet Venus has only recently had a stable orbit. Before this, it moved extensively around the solar system, and specifically passed extremely close to the sun. Because of this, the surface of Venus is hotter than that of earth. That the surface of Venus is hot was, like many of Velikovsky's claims, contrary to mainstream scientific opinion at the time. Nevertheless, Velikovsky turned out to be correct, and mainstream scientists wrong on this particular point

[10] Earlier theories are usually more general in the sense that they consist of fewer theoretical assumptions. Their domain of application, of course, is usually narrower than later theories.

(although the scientific explanation for the surface temperature of Venus of course differs from Velikovsky's). So a direct deductive consequence of Velikovsky's theory was novel, surprising by the standards of contemporary science, and in fact empirically accurate.

Nevertheless, as Vickers argues, this successful novel prediction does not necessarily serve to confirm Velikovsky's theory to any great degree. Firstly, the theory made *many* novel predictions, and that about the surface temperature of Venus is one of only a small number that turned out to be true. The no-miracles argument does not give any grounds for regarding novel prediction as a special source of confirmation for the theory in this sort of case because the alternative explanation, that a lucky coincidence has occurred, is in fact quite plausible. It is, in other words, not surprising at all if some predictions of a false theory turn out to be true with such a 'scattershot' approach.

Thus far, this chapter has largely addressed the subject of accurate empirical predictions, and how these might serve to confirm a theory. Before analysing the Velikovsky example any further, therefore, it is worth making some more general remarks about *inaccurate* predictions, and how these might *disconfirm* a theory. Notice that most accounts of predictive success discussed in this chapter, including the temporal and UN accounts, focus specifically on the relation between a particular empirical result and the theory. This sort of account therefore includes by default the sort of predictive success exemplified by the Velikovsky case as a confirming instance for the theory. However, each of these criteria draws its strength from the no-miracles argument with the result that this sort of case can therefore be excluded by returning to first principles. This is exactly the approach that Vickers takes. In contrast to these accounts, the UV of empirical success invokes a '*global'* criterion. In particular, a theory is confirmed under the UV just in case it is simple relative to the entire body of verified empirical consequences it entails. Therefore, unlike with more 'local' criteria, such as the temporal account, there is no need to resort to first principles to exclude the Velikovsky case under the UV. Velikovsky's theory can be very easily dismissed simply by noting that it is both extraordinarily complex and has very few verified empirical consequences.

Surveying cases of 'empirical failure' more broadly, however, it is clear that this simple response will not suffice in general[11]. Consider, for example, the proposition "All rabbits are white". This proposition is intuitively not at all complex, and there are many instances where it is satisfied. This 'theory', in fact, entails many more verified empirical results than are required to construct it. However, there are many more instances where it is *not* satisfied than where it is. To what extent, then, is the theory disconfirmed by the observation of a single brown rabbit? The simple response that the theory is straightforwardly refuted is not satisfactory. In line with the partial realist thesis articulated in this thesis, one might still want to make room for the claim that the theory 'latches onto' some genuine regularity, albeit imperfectly. It might, in other words, be approximately accurate, accurate under certain circumstances, accurate as a limiting case of a more general theory, etc.

The question then is how one might distinguish hypotheses that are genuinely, if imperfectly, empirically successful from those that are merely lucky. The intuitive response is that a theory can still be considered successful if it does not make 'excessively many' false predictions. The problem with this response, of course, is how to define 'excessive'. In the general spirit of the UV, here follows a tentative suggestion. The thing to notice about the Velikovsky case is not merely that it is complex, but that correcting its many empirical failures would require the addition of many (additional) *ad hoc* posits, thus increasing its complexity even further. Thus, even if the theory could avoid refutation, it would sacrifice any empirical success in doing so. The theory is not empirically successful overall because, again speaking intuitively, its empirical failures 'overwhelm' its empirical successes. Generalising from this point, the empirical success of a theory which makes some inaccurate predictions can be assessed by, first, determining what *ad hoc* hypotheses would need to be added to the theory to save it from refutation. If the theory modified in this way is still empirically successful under the UV account,

[11] This point has been pressed on me my Orri Steffanson,.

especially if the *ad hoc* modifications imply new empirical consequences which turn out to be accurate, the unmodified theory should also be counted as successful. Otherwise, not. Call this the “prospective success criterion”.

This suggestion is admittedly rather rough, but it draws at least intuitive support from a line of thought that arises in both Kuhn (1962/1996) and Lakatos (1968). Both authors emphasise that scientists do not necessarily (or even frequently) abandon a theory/paradigm/research programme which encounters apparent falsifiers. This is because they have confidence that the theory will eventually be able to deal with these problems in due course. This confidence is, presumably, partly underpinned by a sort of induction from the ability of the theory to successfully accommodate similar cases in the past. However, it is plausible that scientists also employ something like the prospective success criterion. They acknowledge that the theory faces potential falsifiers but judge that, whatever modifications to the theory turn out to be necessary in defeating these falsifiers, they won’t be sufficient to undermine its empirical success.

Thus concludes one line of criticism against Velikovsky’s theory. The general suggestion emerging from this criticism is that, for a theory to count as confirmed under the UV, it cannot entail an “excessive” number of falsified empirical results. A separate criticism of Velikovsky’s theory offered by Vickers is that the prediction of the surface temperature of Venus is “vague”. As Vickers points out, just about any temperature hotter than that of Earth would probably be counted as satisfying the description “hot”. So it was in fact not very unlikely for this prediction to come out correct simply by chance. Any attempt to offer the truth of the theory as the best explanation of the accuracy of the prediction is thereby blocked. Vickers therefore suggests as a general rule that a theory be counted as predictively successful only if the predictions it makes are sufficiently “impressive”. Or, as Popper would put it, a theory must “stick its neck out”!

That a theory deductively entails verified empirical results is therefore not sufficient for it to be confirmed under the UV. It cannot entail too many falsified results alongside the verified ones, and it cannot entail the verified results too vaguely. However, it is also not necessary that a theory *entail* verified results to be regarded as confirmed. There are also occasions where a theory is confirmed by merely ‘suggesting’ or ‘inspiring’ certain empirical results. This claim is supported by detailed arguments in Chapter 3.10, in direct opposition to arguments made by Vickers and Cartwright. Therefore only a brief argument for this claim is given, as follows.

Note that, from the perspective of pure logic, there is nothing epistemically distinctive about a relationship of deductive entailment between theory and empirical result. Inferring the truth of the theory from the truth of the empirical result is *always*, strictly speaking, an instance of the fallacy of affirming the consequent. However, it has been accepted (at least for the sake of argument) that it is permissible on certain occasions to infer that a theory is true because that is the best *explanation* of the fact that it gives rise to true empirical predictions. But adopting this sort of explanatory reasoning does not imply any *a priori* stance on what sort of “gives rise” relation is allowable. It is certainly extremely plausible that, all else being equal, the existence of a deductive relationship makes the explanatory argument stronger. But there could nevertheless be various circumstances in which the existence of a somewhat ‘weaker’ relation between theory and verified empirical result is sufficient to regard the former as confirmed. One example is where an abstract theory only “guides” the formation of a successful empirical model (see Chapter 3.4 for a detailed discussion of these cases), but does so for *many* such models. The unifying power of a hypothesis can, in other words, be regarded as the combination of at least two factors, namely the ‘breadth’ of the empirical results that follow from the hypothesis and the ‘strength’ of the logical connection between the hypothesis and these results.

Previously in this chapter, the criterion of unifying power has been characterised as a function simply of the number of verified empirical results which a theory entails and the complexity of this theory. In this section, it has been argued that several other factors must be taken into consideration before a theory can be regarded as confirmed. The first is how many *falsified* empirical results it gives rise to. The second is the *precision* with which these results are predicted. The third is the '*strength'* of the logical connection between theory and results. Putting all this together gives a new version of the UV, as follows:

> A theory is confirmed just in case, and to the extent that, it gives rise to more verified empirical content than that required to construct it, does so with precision, and does not give rise to excessive falsified content. (UV3)

Where "give rise" is intended to indicate that relations that are logically weaker than deductive entailment can be included, and "excessive" should be understood in the context of prospective success criterion proposed above.

As suggested above, there may be some trade-off between the various factors involved in this definition – a theory that, with only very 'natural' assumptions and boundary conditions, deductively entails a small number of accurate predictions may be just as empirically successful as one which very loosely 'guides' the construction of many empirically accurate models. There will be no attempt to give anything resembling an exact or comprehensive treatment of these trade-offs in this thesis. I concede immediately that there will therefore be many cases of theories where it is simply unclear whether or not a theory is sufficiently empirically successful under UV3 to warrant a realist attitude towards it. There will also, however, be cases that fall relatively clearly on one side or the other. Hopefully the ambiguous cases will in fact comprise a small minority of episodes in the history of science. In any case, the examples presented below belong to the subset (however small or large it turns out to be) of relatively clear cases. Or so it will be argued.

## **11. Chapter summary**

I have argued that the NMA demands an account of empirical success which satisfies the intuition that successful theories (or at least the parts of these theories responsible for their success) are true. One popular candidate for such an account holds that empirical success should be understood as novel predictive success. There are, however, competing notions of how "novel prediction" should be interpreted, including the temporal, theoretical and use-novelty accounts. I argue that all of these accounts, because they claim that what theories should be counted as empirically successful depends on historically contingent factors, are not adequate to the NMA.

I have also examined various forms of weak predictivism, which claim that predictive success is not intrinsically a reason to prefer a theory, but merely indicates that some more fundamental epistemic virtue has been satisfied. The severe testing approach, advocated by Howson and especially by Mayo, is a form of weak predictivism which, I have argued, also falls prey to the concern about historical contingency. However, the type of weak predictivism which holds that theoretical unifying power is the fundamental theoretical virtue is considerably more promising.

In fact, I argue that Worrall, on paper an advocate of the use-novelty view, is in fact committed to a particular variant of the theoretical unification idea. I have in consequence dubbed a position derived from his account the unification view (UV), and argued that the UV adequately captures the intuition underlying the NMA. The UV is concisely captured by the slogan: A theory is empirically successful just in case it entails more empirical content than that required to construct it.

When referring to the empirical content required to construct a theory, we do not mean simply that required to fix whatever free parameters are evident on a face-value inspection of the theory. The general framework of the theory embodies substantive assumptions, and establishing these assumptions also involves empirical information. Judging that a theory is empirically successful on an absolute scale therefore involves quantifying, *via* rational reconstruction, the empirical content required to construct it "from nothing".

Finally, I have demonstrated that a relationship of deductive entailment between theory and empirical results is neither sufficient nor necessary for the theory to count as confirmed under the general principles of the UV. I have therefore proposed a more general characterisation of the UV. This states that a theory is confirmed just in case, and to the extent that, it gives rise to more verified empirical content than that required to construct it, does so with precision, and does not give rise to excessive falsified content.

# **Chapter 3. What parts of scientific theories are "essential"?**

## **1. Chapter overview**

Several specific forms of partial realism have been proposed in the literature, including "entity realism" (Hacking, 1983; Cartwright, 1983), "phenomenological realism" (Cartwright, 1999a, 2009; Cartwright *et al.*, 1995; Shomar, 1998; 2008), "structural realism" (Worrall, 1989a, 2007), the "working posits" idea (Kitcher, 1993), the "*divide et impera*" strategy (Psillos, 1999), "Ramsey-sentence realism" (Cruse & Papineau, 2002), and "semi-realism" (Chakravartty, 1998, 2007). Although there are many points of commonality and difference between them, these positions differ most substantially in how they define the vague term "essential" in the definition of partial realism. The aim of this chapter is to produce a satisfactory definition of "essentialness" and thus give an account of what parts of scientific theories we ought to be realists about (if we are to be realists at all).

I begin, in section 2, by examining the direct reference theory of meaning. This has been applied as account of essentialness in its own right, but is more commonly employed as part of a hybrid "causal-descriptive" account. It is argued that a pure direct reference account is manifestly unsatisfactory for partial realism and that, in a hybrid account, only the descriptive component substantially contributes to picking out essential elements.

Section 3 compares retrospective and prospective approaches to essentialness. The former apply knowledge of current theories to pick out the essential elements of past theories, whereas the latter do not. It is argued that, while not totally trivial (as some have argued), retrospective approaches are decidedly less interesting than prospective approaches.

Sections 4-6 provide detailed examinations of several of the particular accounts of essentialness that have been proposed in the literature, namely entity realism, phenomenological realism, structural realism and semi-realism. Ramsey-sentence realism, it is shown, is best understood as a form of structural realism. It is argued in section 7 that the workings posits strategy and the *divide et impera* move are sufficiently similar that they can be treated as a single position. Hence the term “working posits” is used throughout to refer to both accounts. This position is examined at some length, in sections 7-9, in large part because the notion of a “working posit” is so open to interpretation. Section 9 is focused on a particularly sophisticated interpretation of the general position advanced by Vickers.

The account of essentialness advocated in this thesis has not yet been seen in the literature, and is articulated and defended in section 10 of this chapter. The basic idea behind it is that the criterion of unifying power developed in Chapter 2 ought to be applied to *parts* of theories, rather than theories as a whole. The chapter is summarised in section 11.

## **2. Direct reference approaches**

One of the primary concerns motivating referential approaches is the theoretical incommensurability thesis raised by, among others, Kuhn (1962/1996)) and Feyerabend (1975/1993). The concern here is that the meaning of a term is dependent on the larger theory in which it is embedded. So the term “mass”, for instance, has one meaning in Newton’s theory, where it is described as a stable property of objects, and a very different meaning in Einstein’s theory, where it is understood to be dependent on how a given object has been accelerated. The incommensurability thesis can be framed as a problem for scientific realism along the lines of the pessimistic induction. This is because it implies that successful past theories fail to accurately represent the world by the light of our current best theories. Even if these past theories use some of the same theoretical terms present in current theories, the entities designated by these terms do not refer to any entities which actually exist.

Referential approaches attempt to overcome this objection by appealing to the direct reference account of meaning proposed by Putnam (1975b, 1978) and

Kripke (1980). Under this account, the meaning of a term is identified with its *actual* referent, rather than the properties associated with it by some theoretical description. For instance, the meaning of a natural kind term such as “electron” is none other than the actual natural kind which is the extension of the term at the time it is first used, if there is any such extension. The “stereotypes” associated with this entity (i.e. the properties ascribed to it) may differ between theories, but referential continuity is retained between theories, so long as the usage of the term in the new theory is derived from the initial usage. That is, although Rutherford, Bohr and Schrödinger, for instance, may have had different theories of electrons, they were all attempting to describe the same underlying entity.

Significantly, if there is a dispute about the proper reference of some term, the direct reference account enjoins us to examine what particular entity the person or people who dubbed the term were actually causally interacting with at the time they dubbed it. This is the point of the celebrated “twin earth” thought experiment (Putnam, 1973, 1975b). Suppose that there is a planet – twin earth – on which there is a substance which resembles water in all outward aspects, but is actually completely chemical distinct from the substance $H_2O$. Suppose, furthermore, that a human being from earth travels to this planet and there encountered this substance. Because she is completely unable to distinguish it from regular earth water, the traveller refers to this new substance as “water”. She is, according to Putnam’s account, simply mistaken in the use of this term (although understandably). The correct reference of the term “water” is still the particular natural kind with which English-speakers were acquainted when they coined the term, namely that with the chemical formula $H_2O$.

Importantly, the reference of a term is fixed by this sort of causal association even if the association between those who initially use the term and the corresponding referent is mediated by a substantially false description. For instance, late nineteenth-century scientists who began to acquire systematic evidence for the existence of atoms may have understood them as indivisible particles. Yet, although it is now believed that atoms are divisible under certain circumstances, the direct reference theory holds that these scientists were indeed referring to the same underlying natural kind as we are today.

This notion of letting reference follow causal interaction is the basis for a certain realist strategy for picking out the essential parts of theories. Here, for instance, Hardin and Rosenberg discuss the posit of a luminiferous ether in Fresnel’s wave theory of light:

> “Looking back across the range of theories from Fresnel to Einstein, we see a constant causal role being played in all of them; that causal role we now ascribe to the electromagnetic field. One permissible strategy of realists is to let reference follow causal role. It seems not unreasonable, then, for realists to say that ‘ether’ referred to the electromagnetic field all along.” (Hardin & Rosenberg, 1982, pp. 613-614)

And here is Kitcher’s somewhat more detailed argument along the same lines:

> “Consider two hypotheses about the mode of reference of Fresnel's term 'light wave,' or, more exactly, about the tokens of that term that figure in his solutions to problems of interference and diffraction.
>
> *HR.* Fresnel's dominant intention is to talk about light, and the wavelike propagation of light, however that is constituted. He has, of course, a false belief about the medium of propagation. But, since his primary aim is to discuss light and its wavelike qualities, his tokens of 'light wave' in the solutions of the diffraction and interference problems genuinely refer to electromagnetic waves of high frequency.
>
> *HN.* Fresnel's references are explicitly fixed through the descriptions he gives at the beginning of the memoir in characterizing the wave theory of light and in introducing the wave equation. Tokens of 'light wave' that occur later in the memoir – for example in the solutions to the problems of

> diffraction – thus have their references fixed by the description "the oscillations of the molecules of the ether." Since there is no ether, they fail to refer…
>
> There is a lot to be said for *HR.* The sole function of the ether in the prizewinning memoir—and throughout *most* of Fresnel's writings—is to answer to the felt need for a medium in which light waves propagate. Fresnel typically makes no detailed claims about the nature of this medium. Why, then, should we give priority to a description making reference to an entity about which Fresnel would surely have admitted his almost total ignorance, rather than seeing his dominant intention as that of talking about the wavelike features of light, *however they happen to be realized?*" (Kitcher, 1993, pp. 146-147, emphasis in original)

Similar views are also discussed sympathetically by Devitt (1984), Cummiskey (1992), Psillos (1994), and Niiniluoto (1999, pp. 120-132).

So a theoretical term introduced to explain the appearance of a certain phenomenon then refers to whatever entity or entities are actually causally responsible for that phenomenon (and hence for the introduction of the term). The basic difficulty with this sort of account, as pointed out by Laudan (1984), is that practically *any* two theories which attempt to explain a common set of phenomena will therefore refer to the same set of entities. For example, Aristotle's notion of "natural place", Descartes' particle "vortices" and Newton's "action at a distance" all play the causal role of accounting for gravitational action. So, if we were to cash out the notion of "preservation" in terms of similarity of causal role, virtually any two theories in the same domain would trivially satisfy the preservation requirement.

Psillos (1999, ch. 12) acknowledges the force of this argument, and so argues that a simple causal theory of reference is not sufficient for the needs of scientific realism. Instead, he advocates a hybrid "causal descriptivism", under which there is referential continuity between two theories just in case (i) they both posit entities which fulfil the causal role of accounting for a given class of phenomena; *and* (ii) they ascribe some of the same properties to these entities, these properties giving rise to the relevant phenomena within the respective theoretical models. In his words:

> "... although there may well be nothing in the world which possesses all the attributes ascribed to an abandoned posit *a,* there may well be a current posit β to which are ascribed some (sometimes most) of the attributes ascribed to a, and which is also considered to be causally responsible for the same phenomena as *a* had been taken to produce. Should this situation occur, we may be willing to say that the term intended to refer to the abandoned posit a refers (or, at any rate, *approximately refers*) to the current posit β." (Psillos, 1999, p. 284)

While this is certainly a sensible corrective to the naive causal view – even the realist Worrall (1989a) describes Hardin and Rosenberg's treatment as "far-fetched" – a few additional remarks are in order. The crucial point, which was made by Laudan, is that the first condition of Psillos' account will be satisfied by many or most of the examples of interest. All of the work, therefore, in distinguishing cases that do and do not satisfy this account will be done by the second condition. The notion of referential continuity thus becomes, by itself, insubstantial. Those who are so inclined are, of course, perfectly entitled to claim that theoretical terms which exhibit the right sort of descriptive continuity also possess referential continuity. But those who are otherwise inclined are also entitled to ignore the concept of referential continuity altogether. All the substantial debates about whether or not a given posit is to be regarded as essential will take place at the level of descriptive continuity. And the notion of descriptive continuity will be cashed out by an account along the lines of structural realism, Psillos' *divide et impera* strategy, etc. The remainder of this chapter is therefore concerned with determining which, if any, descriptive account along these lines is satisfactory.

## 3. Retrospective versus prospective accounts of essentialness

In the previous section, it was argued that ‘pure’ referential approaches are insubstantial or vacuous. That is, they are satisfied by just about any two theories that attempt to account for the same empirical phenomena. To avoid vacuousness, it was argued that an acceptable account of essentialness must focus on the properties or relations attributed to the posited entities by a theory. However, before considering any particular account of essentialness along these lines, it is worth examining another, more general, accusation of vacuousness that has been levelled against partial realism. The most detailed version of this accusation is due to Stanford. He argues that Psillos’ and Kitcher’s versions of partial realism are ultimately trivial because they make only *retrospective* judgements about which elements of a theory are responsible for its successes:

> “... [O]ne and the same present theory is used both as the standard to which components of a past theory must correspond in order to be judged true and to decide which of that theory's features or components enabled it to be successful. With this strategy of analysis, an impressive retrospective convergence between judgments of the sources of a past theory's success and the things it "got right" about the world is virtually guaranteed: It is the very fact that some features of a past theory survive in our present account of nature that leads the realist both to regard them as true and to believe that they were the sources of the rejected theory's success or effectiveness.” (Stanford, 2003b, p. 914)

This argument is repeated by Stanford in his (2006, pp. 166-168) and a similar complaint is levelled by Newman (2010, pp. 416-417). The crucial issue is that an “essential” element of theory is defined simply as “that which we now happen to know was preserved in successor theories”.

There are three distinct questions that arise from Stanford’s argument. The first is purely exegetical, namely whether Psillos and Kitcher (and the various other authors implicitly included in this criticism) are in fact making purely retrospective judgements of essentialness. The second question is whether or not this manoeuvre of retrospective judgement does in fact render partial realism a totally trivial position. Suppose that the strong form of Stanford’s objection can be avoided, and that backwards-looking partial realism is not trivial. In this case, a third question is whether there might be other reasons for avoiding retrospective judgements of essentialness.

Regarding the first question, there certainly are certain respects in which Psillos and Kitcher can reasonably be interpreted as offering retrospective accounts be. Both authors appeal to a causal theory of reference to designate which theoretical elements are preserved in episodes of theory change (as discussed in section 2). And, since judgements about the extension of a given theoretical term are generally made from the vantage point of our current theoretical ontology, these judgements are inherently backward-looking. Both accounts, however, also have a descriptive dimension, and they in fact both advocate variations of the “working posits” view (see sections 7-9 of this chapter). Although it will depend on which particular definition of a working posit is proposed, this kind of view can, in principle, be applied equally well retrospectively and prospectively. Moreover, as demonstrated in section 7, neither of them think of themselves as offering a retrospective account.

The second question is whether all accounts of essentialness which rely on examining past theories in the light of current theories are indeed trivial or insubstantial. To put this another way, is it in fact “virtually guaranteed” that these accounts will judge as essential precisely those posits of past theories which happen to be preserved in current theories? The answer provided here is that these accounts are not trivial, because they are not interested simply in whether claims regarded as true by current theories are, as a matter of brute fact endorsed by past theories, but whether the truth of these claims *explains* the success of these past theories.

A good way into this issue is by returning to Post's generalized correspondence principle: "... [A]ny acceptable new theory L should *account for* the success of its predecessor S by 'degenerating' into that theory under those conditions under which S has been well confirmed by tests." (Post, 1971, p. 228, emphasis added). While Post viewed this as a heuristic for eliminating potential new theories, suppose that we instead assume that some independent criterion of empirical success is overriding in judging whether or not a new theory is to be accepted. In this case, the correspondence principle can be understood as a device for identifying the essential elements of the past successful theory. First, we check whether the new theory can account for/explain the successes of the past theory. Then, if it can, we determine which elements of the past theory we appeal to in providing this explanation. These elements are, by this method of accounting, the "working posits" of the older theory.

To avoid the charge of triviality, therefore, all that is required is the possible existence of an older theory whose success cannot be explained in terms of the categories offered by the newer theory. The mere fact that the older theory posits the existence of entities also posited by the newer theory is not sufficient to avoid a refutation of partial realism. For this sort of case to satisfy the variant of partial realism under discussion, the older theory must also attribute properties to these entities in such a way that it is clear, from the perspective of the newer theory, *why* it was empirically successful. Moreover, the ongoing debates around particular case studies seem to address exactly this question. For instance, one contentious issue in the debate around Fresnel's wave theory of light is the extent to which the derivation of the important results of the theory appealed to the *mechanical* properties, such as elasticity, that were attributed to the luminiferous ether. And, if it is decided that these mechanical properties were "working posits" in this derivation, it is *prima facie* difficult to see how to account for the success of the theory, since current theories of light do not attribute any mechanical properties to the electromagnetic field. So, in answer to the second question, it seems that a backwards-looking variant of partial realism is not totally trivial.

The third question can be rephrased as: “Is a retrospective variant of partial realism importantly *less interesting* than a prospective variant?” The answer offered to this question here is “yes”. The generalized correspondence principle can be viewed as either a positive or a negative heuristic. Used as a negative heuristic, it tells us to discard any newly mooted theories which are unable to explain the successes of past theories. Used as a positive heuristic, on the other hand, it tells us which elements of the old theory ought to be used in the construction of any new theory. It seems clear that Post intended his correspondence principle as a negative heuristic. All realists, moreover, must accept *at least* this negative version of the principle. In the remainder of this section, however, it is argued that realists should *also* favour a positive heuristic and thus a *prospective* account of which theoretical elements are essential.

One major argument for the positive heuristic is that the distinction between it and the negative heuristic is untenable in the first place. In the paradigmatic applications of the (negative version of the) correspondence principle, the newer theory is not simply required to replicate the brute empirical results of the older theory, but often relatively abstract theoretical elements. For instance, to use a common example, when Einstein formulated his special theory of relativity, he checked that his equations of motion approximately reduced to the equivalent Newtonian equations under low-velocity conditions. Note that this represents a belief on his part that the Newtonian equations would continue to accurately describe low-velocity scenarios in general, including an infinite number that had not yet been empirically observed by anyone. However, once it is accepted that certain higher-level theoretical posits should be counted as “successes” of the older theory and must be “accounted for” by the newer theory, then there is a need for some principles for deciding *which* of the many theoretical posits on offer count as successes. Einstein, for instance, judged (and rightly so) that Newton’s equations of motion were successful elements of the theory that ought to be preserved in some form in his newer theory, whereas Newton’s assumption of absolute space and time were not. So, to the extent that the realist accepts application of the correspondence principle at all, she also tacitly accepts the application of forward-looking criteria for which elements are essential to the success of a theory.

Another argument for a prospective view of essentialness is simply that it is logically stronger than a retrospective view. The retrospective view has a certain built-in flexibility; it simply claims that we will *somehow* be able to account for the success of past scientific theories from the perspective of our current best theories. A prospective view, on the other hand, is perforce committed to some *specific account* of theoretical essentialness. On broadly Popperian grounds, prospective accounts therefore ought to be favoured in the first instance, simply because they are more falsifiable. This is not to claim that the partial realist cannot legitimately retain the "fall-back option" of a more modest retrospective account. But this is not an option that should be exercised until all hope for a prospective account is forlorn. And there is nothing to indicate that this is currently the case.

A final argument for the positive heuristic (and so a prospective account of essentialness) is that scientists tacitly apply such a heuristic in any case. As mentioned, at least by the beginning of the twentieth-century, physicists had a reasonably good idea about which parts of Newton's theory were worth keeping and which were not. And anyone with scientific training will have similar intuitions about their own field of expertise. We would, for instance, be hard-pressed to imagine a future science of chemistry which lacked the concept of chemical atoms; though there are many other concepts in chemistry to which our allegiance is considerably less secure. Because scientists already do employ a positive heuristic at least intuitively, and this is a significant force in scientific theorising, it is worth attempting to characterise it explicitly and rigorously. And it is also worth attempting, if possible, to defend (or attack) the heuristic so uncovered.

The opponent of prospective judgements of essentialness will, however, surely argue that the line of thought sketched out above is too ambitious. This point is addressed particularly by Stanford in his response to reviewers of *Exceeding our Grasp.* He asks, firstly, if it is indeed possible to identify prospectively which elements of theories are essential to their empirical successes, "... why did we (or the relevant scientific communities) ever believe more than those parts or aspects of past theories on which their empirical successes really depended?" (Stanford in Saatsi *et al.*, 2009, p. 385) Secondly, in response to the observation that scientists themselves frequently make judgements concerning which parts of their theories are essential, he points out that scientists "are routinely *mistaken* in central cases" (*ibid.*, p. 386, emphasis in original). Finally, we can make the additional objection that, if we were able *in general* to identify which elements of scientific theories are essential to their success without the benefit of hindsight, we could perform the same trick upon our best contemporary theories. And then *we philosophers* would be able to do the hard work of science ourselves and simply say what the next, even more successful, theory will look like. Yet this seems deeply implausible (both Stanford and Saatsi make this point in Saatsi *et al.*, 2009).

These objections are mistaken on several counts. Taking the last objection first, it should be emphasised that the task is not to *do* the work of science, but rather to characterise, articulate in more rigorous logical terms and (if it turns out to be defensible) defend a practice which is already present in science. In this respect, the task at hand resembles that of philosophers working on confirmation theory, those attempting to understand the notion of causality, and so on. Of course, if some part of this philosophical commentary allows scientists to apply their existing methods more rigorously and so feeds back into practice, so much the better. But it is highly unlikely that the very general positive heuristic which articulated here will by itself improve upon the judgement of scientists who are deeply immersed in a particular theoretical framework and experimental or observational practice.

It is also certainly not plausible to suggest that the mooted positive heuristic *by itself* would ever tell us what the improved successor to a given scientific theory will look like. In most of the cases examined in this literature, it is taken for granted that only a small proportion of the theoretical posits of even a highly successful theory will typically be viewed as essential to this success. Moreover, as discussed in Chapter 1.6, these posits will often be preserved only "approximately" in the successor theory, and there is a variety of quite different cases that can reasonably be counted as instances of “approximate preservation”. So any sensible positive heuristic would leave some considerable scope for creative redevelopment of the essential posits. Finally, it is frequently the case that the phenomena successfully described by an older theory come to be viewed as special instances of a far larger class of phenomena described by a newer theory. So, although the positive heuristic requires that the essential posits of the older theory be (approximately) preserved in the newer theory, it does not specify anything about the larger ‘superstructure’ that is erected around these posits to form the more general successor theory.

Supposing that a prospective account of essentialness is tenable, Stanford professes puzzlement that scientists would *ever* believe in more than the essential parts of theories. In response to this, firstly, it is worth pointing out scientists’ judgements of which elements of a theory are essential to its empirical success will (and should!) depend on the actual empirical successes and failures the theory has accrued. For instance, as emphasised above, Einstein’s view of what was essential to Newton’s theory was a product of a particular time period. By the early twentieth century, it had become quite clear which elements of the overall Newtonian picture were broadly responsible for accurate results, which were inessential and which in fact largely led to *inaccurate* predictions. Any account of partial realism which attempted to anticipate scientists’ own empirically-informed judgements about what was essential to a theory’s success would certainly be an example of overreach, and any should consequently be rejected.

Stanford’s other concern is that, even if we adopt the general policy of deferring to scientists’ own judgements, these judgements are often simply mistaken. This concern is mitigated by the point made above, that these judgements become steadily more trustworthy as more directly relevant empirical information is accumulated. But, Stanford might argue, this response surely predicts that scientists will be more cautious in their judgements of essentialness when the theory in question is still young and relatively few empirical data are on the table. In fact the opposite often seems to be the case – witness the enthusiastic remarks (a beloved source of quotations for anti-realists) made in support of the luminiferous ether posit when Fresnel's theory was young. Here, for instance, is James Clerk Maxwell:

> "Whatever difficulties we may have in forming a consistent idea of the constitution of the aether, there can be no doubt that the interplanetary and interstellar spaces are not empty, but are occupied by a material substance or body, which is certainly the largest, and probably the most uniform body of which we have any knowledge." (Maxwell, 1878)

Similarly, here is Helmholtz:

> "There can no longer be any doubt that light-waves consist of electric vibrations in the all -pervading ether, and that the latter possesses the properties of an insulator and a magnetic medium." (von Helmholtz, 1899)

Several additional examples along these lines are provided by (Cordero, 2011).

Worrall (1990), however, points out that certain scientists in this period were more cautious. He provides a quote by Airy that is apparently just as triumphalist as those given above:

> "The Undulatory Theory of Optics is presented to the reader as having the same claims to his attention as the Theory of Gravitation: namely that it is certainly true. With the regard to the evidence for this theory; if the simplicity of a hypothesis; which explains with accuracy a vast variety of phenomena of the most complicated kind; can be considered a proof of its correctness; I believe there is no physical theory so firmly established as the theory in question." (Airy, 1831/1842, pp. v-vi)

This is followed, however, by:

> “This character of certainty I conceive to belong only to what may be called the geometrical part of the theory: the hypothesis, namely, that light consists of undulations depending on transversal vibrations, and that these travel with certain velocities in different media according to the laws here explained. The mechanical part of the theory, as the suppositions relative to the constitution of the ether, the computation of the reflected and refracted rays, &c., though generally probable, I conceive to be far from certain.” (ibid, p. vi)

This second quote is more in line with the sort of partial realism endorsed by modern scientific realists, and indeed by the structural realism advocated by Worrall himself.

So some contemporary scientists made a full-fledged realist commitment to the existence of a luminiferous ether, whereas others were committed only to the (we now think) more defensible claim that light has certain “geometrical” properties. Maxwell and the other more metaphysically ambitious scientists held that an adequate theory must render *intelligible* the physical system in question, and that intelligibility is only conveyed by means of a mechanical model. Under such a criterion, the posit of a mechanical luminiferous ether is therefore essential to the empirical success of Fresnel’s theory simply because that success is unintelligible without it (see Cordero, 2011 for a more detailed discussion). Airy and other scientists in the metaphysically cautious camp, in contrast, applied some criterion of essentialness that that was more narrowly focused on the logical relationship between particular theoretical posits and the successful empirical predictions and explanations of the theory. Notice, incidentally, that Airy, at least, understands this logical relationship broadly along the lines of the unifying power criterion discussed in the previous chapter. As such, he anticipates the view of essentialness that will be advocated in the final sections of this chapter.

Stanford, surely, is correct that some scientists – including brilliant and influential scientists – are frequently mistaken about which elements of their contemporary theories will come to regarded as respectable by their successors. It is not clear, however, that this undermines the broader aim of establishing prospective criteria for which theoretical elements ought to be regarded as essential. The partial realist is not, after all, committed to the claim that scientists always get things right. On the contrary, it can only serve to further support a particular criterion of essentialness if history tells us both that scientists who have applied this criterion have tended to judge correctly regarding which theoretical elements are likely to be preserved and that those who failed to apply this criterion have tended to judge poorly. The mere fact that scientists have been mistaken therefore does not undermine the possibility for at least certain types of prospective accounts of essentialness.

In this section, it has been argued that, while not trivial, a purely retrospective view of what is essential to the success of a scientific theory is uninteresting. And it is uninteresting not simply because we ought, in a Popperian mindset, to prefer strong theses until they are falsified. It is uninteresting because it fails to engage with a key feature of scientific practice, one that cries out for philosophical articulation. Accounts of essentialness that depend upon hindsight should therefore be rejected. Any satisfactory account must, instead, depend on ‘timeless’ factors such the internal logic of the theory itself and the empirical consequences that follow from it.

## 4. Entity realism and phenomenological realism

Entity realism was proposed by two leading lights of the “Stanford School” of philosophy of science, namely Hacking (1982; 1983; 1988; 1989) and Cartwright (1983, ch. 5). Their stated grounds for adopting this position, however, are distinct. Hacking rejects explanationist arguments for scientific realism and opts instead to pursue arguments that rely on the notion of intervention. His rationale for this is his acceptance of the argument that inference to the best explanation is simply circular when applied to the truth of scientific theories (this argument is given in more detail by Laudan, 1981; and Fine, 1984; see also Chapter 1.3 of this thesis). As he explains his reasoning:

> “Realism and anti-realism scurry about, trying to latch on to something in the nature of representation that will vanquish the other. There is nothing there. That is why I turn from representing to intervening.” (Hacking, 1983, p. 145)

Hacking’s interventionist account of scientific realism is built squarely upon the Kripke-Putnam causal theory of reference, discussed in section 2. However, he does not simply make the uninformative claim that a theoretical term refers to whichever entity the first user of the term in fact causally interacts with. Indeed, he is not concerned with mere causal interaction in itself. Rather, he thinks that we have good grounds for believing in the existence of an entity just in case we have a sufficient understanding of its causal powers to *manipulate* it for certain pragmatic ends. This criterion of realism is captured by his famous slogan: “[I]f you can spray them [electrons], then they are real” (*ibid.*, p. 23). In more detail:

> “The “direct” proof of electrons and the like is our ability to manipulate them using well understood low-level causal properties. I do not of course claim that “reality” is constituted by human manipulability. We can, however, call something real, in the sense in which it matters to scientific realism, only when we understand quite well what its causal properties are. The best evidence for this kind of understanding is that we can set out, from scratch, to build machines that will work reliably, taking advantage of this or that causal nexus. Hence, engineering, not theorizing, is the proof of scientific realism about entities.” (Hacking, 1982, p. 86)

There are two main objections to Hacking’s argument. These are not knock-down objections to entity realism as such, but rather show the need for some additional distinctions and arguments which Hacking unfortunately does not offer. The first objection is that Hacking has not succeeded in eliminating all reliance on explanation (Reiner & Pierson, 1995; Resnik, 1994).

> “Laboratory skills do not give us access to otherwise unobservable entities, but only to certain observable interactions in the apparatus. Only by IBE can we come to believe that these observable signs indicate the presence of causal interactions, that these interactions are not artifacts, and that entities lie behind them. Moreover, only by an additional IBE can we take the evidence as warranting belief in the existence of exactly one kind of entity, rather than two kinds or a thousand kinds.” (Reiner & Pierson, 1995), p. 67)

These authors do not commit either way on the applicability of inference to the best explanation in resolving these sorts of questions. What they do argue is that, if we accept Hacking’s arguments for entity realism, we have no principled grounds for rejecting similarly explanatory arguments for realism about other theoretical posits.

The second, related, objection is that the successful manipulation of unobservable entities such as electrons surely presupposes more than just a naked assertion of their existence. As Psillos puts it:

> “... [I]f it is not admitted that some theoretical descriptions of the causal powers of the entity are correct, then the mere positing of the entity cannot produce any sound expectations about which phenomena are due if and when this entity is manipulated....” (Psillos, 1999, p. 248)

In other words, to assert the existence of an entity, one must suppose the truth of at least those theoretical claims which give a basic description of how the given entity will behave within a certain type of experimental setup. This point is pre-emptively conceded by Hacking, when he claims that we presuppose knowledge of “a modest number of home truths about electrons” (Hacking, 1983, p. 265) when we design scientific apparatus relying upon manipulation of these entities. Hacking does not, however, give any sort of general account regarding which purported properties of entities are to be identified as the “home truths”.

Cartwright’s argument for entity realism evades both of the objections given above (the differences between the two forms of entity realism are discussed by Clarke, 2001). An exposition of her views therefore stands in place of a direct response to these objections. Cartwright proposes a form of argument she terms “inference to the most probable cause”, which is recognisably and explicitly a form of IBE. As such, she effectively concedes the point made by Reiner and Pierson. She does, however, explicitly draw a distinction between “causal explanation” and “theoretical explanation”:

> “What is special about explanation by theoretical entity is that it is causal explanation, and existence is an internal characteristic of causal claims. There is nothing similar for theoretical laws... God may tell you that Wollheim's paper is after mine, and that his paper is true. You have no doubts about either of those propositions. This signifies nothing about the truth of my paper. Similarly, God tells you that Schroedinger's equation provides a completely satisfactory derivation of the phenomenological law of radioactive decay. You have no doubt that the derivation is correct. But you still have no reason to believe in Schroedinger's equation. On the other hand, if God tells you ... that the ionization produced by the negative charge explains the track in the cloud chamber, then you do have reason, conclusive reason, to believe ... that there is an electron in the chamber.” (Cartwright, 1983, p. 93)

So theoretical explanations still work *as explanations* if the more general law found in the explanans is false. In contrast, a causal explanation simply fails as an explanation if the posited entity does not exist. To be committed to a causal explanation is also to be committed to a certain *existential assertion*. This is, as it were, a ‘transcendental argument’ for drawing a distinction between causal and theoretical explanations. It starts from the assumption that we do have some successful explanations, then reasons backwards to see what conditions must be in place in each case for this to be so.

Cartwright also indirectly addresses the second objection to Hacking's entity realism given above. She concedes the point that, if we do successfully manipulate entities and offer causal explanations, *some* of our theoretical assertions must be true. However, she thinks that there is a distinction between asserting that entities possess certain causal powers and asserting that they obey abstract theoretical laws. This connects to the titular thesis of her book, that the laws of physics "lie" (see especially the introduction and ch. 2 of this book). She does not claim that *all* the laws of physics are false. Rather, she distinguishes "phenomenological laws" from "fundamental laws", and expresses a generalised scepticism about only the latter. A phenomenological law is one that attempts to *describe* some particular experimental situation, and Cartwright thinks that many such laws are essentially accurate. Causal explanation is respectable precisely because it is the type of explanation involved in descriptively accurate phenomenological laws. To predict and explain the behaviour of some particular system requires that we understand the relevant causal powers of the entities involved in that system.

Fundamental laws, in contrast, attempt to *explain* a wide variety of particular phenomenological laws. These, however, do not accurately describe all, or even most, actual states of affairs in their purported domains. They can, on the one hand, be interpreted as accurate descriptions of a very limited number of situations: experiments conducted under tightly controlled conditions, for instance. On the other hand, they can be interpreted as "*ceteris paribus* laws". *Ceteris paribus* laws tell us "which factors are explanatorily relevant" (*ibid.,* p. 48), but they do not deductively predict what will happen; they will always be subject to exceptions or competing explanatory factors. So, insofar as they are taken as literal descriptions of events, fundamental laws are false.

Before addressing criticisms of Cartwright's version of entity realism, I will examine how it has evolved into so-called "phenomenological realism" in the work of

Cartwright and her students. In her more recent work, Cartwright has focussed more on models than on laws. In particular, she has endorsed the thesis that models represent and explain physical systems independently of theory (Cartwright, 1999a; 1999b; see also Morrison, 1998; 1999). This "autonomy" thesis ultimately depends on a thesis about how models are constructed. The "received view" of the relationship between theory and model supposes that a model of a particular physical system is the deductive consequence of some theory. In fact, this is only true of "theoretical models". "Phenomenological models", in contrast, are "built upwards" by attempting to fit the phenomena directly (a very similar view, although in slightly different terms, is also advocated by Cartwright (1983, ch. 8). And, argue Cartwright and Morrison, phenomenological rather than theoretical models are typical of successful scientific practice (this distinction is obviously connected to that between phenomenological and theoretical laws sketched above). The primary example used by Cartwright and her co-workers to illustrate this claim is the empirically extremely successful London model of superconductivity (Cartwright *et al.*, 1995; Suarez, 1999). Cartwright *et al.* argue that there was no theoretical rationale for this model in the theory of electromagnetism. Rather, the model was "built upwards" specifically to accommodate the observed phenomenon that superconducting materials rapidly expel any externally applied magnetic field as they undergo the transition to superconductivity. Fundamental laws are not idle in this process, but their role is to provide heuristics which "guide" the formation of the phenomenological model.

Although Cartwright ( 2009) has discussed it favourably, the main explicit advocate of "phenomenological realism" is Shomar (1998; 2008; 2009). Phenomenological realism endorses an instrumentalist attitude for abstract theory but realism for phenomenological models. As expressed by Shomar, the major motivation for this bifurcated position is as follows: "[Phenomenological] models are closely related to the empirical findings. Therefore, it is more probable that what they say about nature is correct" (Shomar, 2009, p. 324). Notice that the emphasis here is subtly different than that of Cartwright's approach in *How the Laws of Physics Lie*. In this earlier work, we are enjoined to be realists about entities and their causal powers, whereas now we are enjoined to accept the truth of certain (albeit low-level) theoretical propositions. I take it, however, that the crucial issues for the realist are that in both presentations (i) we are enjoined to accept the existence of entities which figure in successful phenomenological laws, and to accept attributions of causal powers to these entities; and (ii) the justification offered for accepting these propositions is that they are the best explanation of the success of these phenomenological laws. For simplicity, therefore, I will elide differences between the earlier and later views and simply refer to “phenomenological realism” in what follows.

This brings us to some criticisms of phenomenological realism. One thing I should make clear is that more ambitious forms of scientific realism are committed to *at least* phenomenological realism. Shomar is certainly correct in saying that phenomenological models are more directly related to empirical findings than abstract fundamental laws. As such, if *any* part of a scientific theory can be inferred to accurately represent the world by way of explanationist arguments like the NMA, then surely phenomenological models are first in line. The question, however, is whether we can justify being realists *only* in respect of phenomenological models. Recall that Cartwright and colleagues give three distinct arguments for the claim that phenomenological laws/models enjoy a qualitatively distinct epistemic status compared to abstract or fundamental laws. The first is that phenomenological laws in general provide causal explanations, and that causal explanations require the truth of the explanans in a way that theoretical explanations do not. The second argument is that, whereas phenomenological laws often accurately describe actual experimental states of affairs, fundamental laws are generally *false* if interpreted as literal descriptions of events. The third argument is that, whereas phenomenological laws are constructed in such a way that they describe phenomena directly, fundamental laws are used much more loosely and heuristically to construct phenomenological laws or models. I will address these arguments in turn, although there will be some overlap between the responses.

Lipton ( 2004, p. 199) addresses himself to the first, ‘transcendental’, argument for distinguishing causal and theoretical explanations. Lipton agrees that a putative causal explanation is no explanation at all if the entity it posits does not exist. “Explain” in this context is, to use Ryle’s (1949) term, a success or achievement verb. The proper use of such a verb requires not only that some action is performed, but that this action achieves a specified end. To take another example, we cannot say a person *knows* some proposition unless that proposition is in fact true, though they could still *believe* it if it were false. So if “explain” in the context of a theoretical explanation is a success verb, a putative theoretical explanation fails to explain if some or all of the propositions in the explanans are false. This, it should be noted, is hardly a novel claim on Lipton’s part. One of the adequacy conditions in the venerable deductive-nomological account of explanation is that “[t]he sentences constituting the explanans must be true” (Hempel & Oppenheim, 1948), p. 137). This is not to say that “explain” *has* to be understood as a success verb; it can also reasonably be interpreted as not requiring any achievement at all, along the lines of “belief”. Indeed, this is exactly how the anti-realist understands the term, for both causal and theoretical explanations. The point is that consistency would seem to demand that we go either one way, into full-fledged realism, or the other, into anti-realism. In this light, Cartwright’s transcendental argument seems like a rather unsatisfactory attempt to draw a distinction by mere fiat.

This reasoning would seem to pose a problem for Cartwright. If Lipton is correct, then a successful explanation requires that the explanans be true. But Cartwright’s second argument, that fundamental laws are generally false, would then seem to imply that there are no genuine explanations that appeal to fundamental laws. Cartwright pre-empts this problem by removing herself from the deductive-nomological framework of explanation. It is only when they are read as “covering laws” which are intended to deductively entail the existence of various phenomena that fundamental laws are false. However, they do a perfectly adequate job in “specifying which factors are explanatorily relevant”. In *The Dappled World*, Cartwright (1999b) extends this approach and makes it more explicit by arguing that fundamental laws tell us about the causal powers or “capacities” of various entities in the world. Notice, then, that Cartwright's second and third arguments offer two sides of the same coin. The second argument tells us what fundamental laws cannot (or at least do not) do, namely deductively entail phenomenological laws. They are not, in other words, covering laws. The third argument tells us what they *can* do, namely tell us enough about causal powers to be helpful in constructing phenomenological laws.

One objection to this picture of the role of fundamental or high-level laws is simply that it is not generally true. Fundamental laws may be related to phenomenological models in various ways. Indeed, we can picture a graduated spectrum, where high-level theory proceeds from more to less directly involved in the success of phenomenological predictions. At one end of the spectrum, a theoretical model that is deductively derived from the fundamental law (subject to the addition of 'natural' boundary conditions, auxiliary assumptions, etc) has the correct form to serve as a phenomenological model. Empirical measurement is required only to fix the values of free parameters left open in the theoretical derivation. In the middle of the spectrum are the cases discussed by Cartwright *et al.* Here both the form of the model and the value of the parameters are derived from the phenomena. Fundamental laws merely provide some helpful heuristics that "guide" this process. At the far end of the spectrum, the phenomena are sufficiently simple in structure that an appropriate model can be read directly from them. High-level theory, if it is involved at all in this sort of case, may serve merely to provide a *post hoc* 'explanation' of the observed regularity.

Cartwright *et al.* are illicitly attempting to base a general view of essentialness on a specific class of cases, including the London model, which lies in the 'middle' of the spectrum described above. Their view therefore fails to capture cases where theoretical models are empirically successful. For instance, an empirically accurate model of the white spot experiment in Fresnel's wave theory is obtainable as a theoretical model from the abstract wave equations and a statement of the boundary conditions. These same wave equations, moreover, can be used to derive predictive models for various other cases. While the wave equations do not *represent* particular experimental circumstances, they nevertheless deductively entail these representations. This is not to argue that the interpretation of the superconductivity case offered by Cartwright *et al.* is incorrect, merely that a satisfactory account of essentialness must be able to capture (at least) *both* types of case.

Finally, it is worth noting that if, following the Cartwright of *The Dappled World*, we interpret high-level theoretical posits as telling us about causal powers, then some sort of realist commitment in respect of these causal powers is not necessarily unwarranted even for cases such as the London model. Indeed, such a realist commitment may be warranted by the sort of “inference to the most probable cause” advocated by Cartwright in *How the Laws of Physics Lie*. The difference is that, in the earlier book, this form of reasoning was implicitly directed only at particular *instances* of causal interaction – the successful manipulation of a track in a cloud chamber licences an inference to the existence of a particular particle which causes the track. Now, in contrast, the successful application of a particular theoretical posit in constructing phenomenological models may be supposed to licence an inference to the existence of a particular *type* of causal interaction implied by this posit. Note, however, that even if this type of inference is warranted, it does not give us reason to believe the truth of propositions, along traditional realist lines. It rather gives us reason to believe in the existence of certain causal powers ‘associated with’ these propositions. But that may still be enough for Cartwright’s more metaphysically ambitious opponents.

## **5. Structural realism**

There have been many advocates of structuralism in philosophy of science and epistemology over the years. There was a distinct burst of activity in this area in the early twentieth century, driven by the then-recent advances in physics towards relativity theory and quantum mechanics, and in formal logic.  Advocates of structuralism in this period, although they did not always describe themselves in these terms, included working physicists such as Poincaré (1905/1952), Duhem

(1906/1954) and Eddington (1939), as well as philosophers such as (Cassirer, 1910/1953), Schlick (1918/1974), Russell (1919; 1927/1992) and Carnap (1928/1967). Recent historically focussed articles by Gower (2000), French (2003) and Frigg and Votsis (2011) provide additional details about these authors. Maxwell (1970a, 1970b) has more recently advocated structural realism, focussing particularly on a development of Russell's view.

It is worth noting, however, that the motivation of many (though not all) of these early structuralists differs quite distinctly from that of modern structural realists following Worrall. These structuralists draw a Kantian distinction between the world of phenomena or percepts and the objective world of "things in themselves". Russell, for instance, understands his task quite explicitly as one of *inferring* something about the structure of this noumenal or objective world from the observed structure of the phenomenal world, on the assumption that the latter is causally dependent on the former. As such, he starts from an initially sceptical empiricist position, and then attempts to develop a more metaphysically ambitious structural realism. For this reason, Psillos (2001) dubs this the "upward path" to structural realism. In contrast, Worrall's structural realism, like other forms of partial realism, is motivated in the first instance by the intrinsically metaphysically expansionary no-miracles argument, and his position is only weakened in response to counterarguments like the pessimistic meta-induction. Psillos therefore dubs this the "downward path" to structural realism.

Despite the illustrious history of the upwards path to structural realism, in this section I shall focus on the recent presentation of the downward path due to Worrall (1989a; 1994; 2007; Worrall & Zahar, 2001), as this approach is currently more influential. This is no doubt due partly to Worrall's (1989a; 1989b) detailed examination of Fresnel's wave theory of light as a case study for scientific realism. Although this example was discussed briefly as a problem case for scientific realism by, for instance Laudan (1981), Worrall's presentation is more detailed and clearly brings out the philosophically relevant features. As such, it has become standard to use this case study as an illustrative example for any novel variant of partial realism. For simplicity and for ease of comparison between the various accounts, I will therefore continue to use this case study as the primary case study in the remainder of this chapter. The broad outlines of this case are given in Chapter 2.3. Here, however, I will give a few more details in order to illustrate precisely why this episode presents a challenge for the realist, before introducing Worrall's structural realism in the context of his own response to the challenge.

On Worrall's interpretation, Fresnel's achievements were two-fold. Firstly, he produced a mathematically rigorous wave theory of light which yielded quantitative predictions for diffraction phenomena. Most spectacularly, it predicted hitherto-unobserved diffraction effects, such as the appearance of a "white spot" at the centre of a shadow cast by a circular disc. Secondly, along with Arago, Fresnel suggested that light was specifically a transverse, as opposed to longitudinal, wave. This formulation accounts for the selective transmission of light through polarising materials. It also resulted in the derivation of the so-called Fresnel equations, which accurately predict the amount of light that will be refracted or reflected at the interface of two media.

This episode presents a *prima facie* problem for the realist because the propagation of a wave appeared, to Fresnel and all other contemporary wave theorists, to require the existence of some medium which permeates space – the "luminiferous ether". Positing the existence of this substance, however, turned out to be quite incorrect from the perspective of later science; in the mature formulation of Maxwell's theory, light is the manifestation of a "freestanding" electromagnetic wave. The standard partial realist response to this case, as established by Worrall, is to attempt to demonstrate that the appeal to a mechanical ether was in fact inessential to the theory's success, but that the abstract wave equations (which were largely preserved in later theories) *were* essential. However, the specific rationale for this manoeuvre differs between the variants of partial realism, as will become evident from the remainder of this chapter.

Firstly, however, let us examine Worrall's own response, which he derives from his own reading of Poincaré:

> "Although Fresnel was quite wrong about *what* oscillates, he was, from this later point of view, right, not just about the optical phenomena, but right also that these phenomena depend on the oscillations of something or other at right angles to the light. Thus if we restrict ourselves to the level of mathematical equations ... there is in fact complete continuity between Fresnel's and Maxwell's theories." (Worrall, 1989a, pp. 118-119)

We should, in other words, not be realists about the intrinsic character of the entities postulated by theory, but rather about the form of the relations between these entities.

There are two immediate objections to this view. The first, due to Psillos (1995; 1999) argues that structural realism simply collapses into a more expansive (though still partial) scientific realism. Psillos begins by calling into question the distinction Worrall seeks to draw between formulating the equations describing the behaviour of an entity and coming to an understanding of what the entity *is*. This distinction is, for Psillos, completely obscure, and is possibly equivalent to the justifiably discarded distinction between "form" and "substance" found in medieval metaphysics. He argues, in opposition to any such view, that to give a structural description of an entity is just the same as giving more information about what the "nature" of that entity. For instance, he argues that "mass" in Newtonian physics is understood as just that entity which, among other things, obeys the second law of motion. Moreover, Psillos points out, it is not only equations which are preserved across theory change. In fact, in the transition from Fresnel's theory to Maxwell's, there is also continuity in various "theoretical principles" such as the conservation of energy and the finiteness of the velocity of light waves. So it is not clear what is special about mathematical equations. Surely, Psillos suggests, it makes more sense to argue simply that Fresnel was right about some theoretical posits, and wrong about others, without committing in advance to these being mathematical equations or some other kind of posit.

The second, related, objection is as follows. If mathematical equations are the *only* thing capable of expressing the structure of a theory, then we have no grounds for realism about theories without any significant mathematical content (Gower, 2000, p. 74; Newman, 2004, pp. 1377-1378). This seems difficult to square with the undeniable predictive successes of, for instance, theories in modern molecular biology.

These two objections together cast significant doubt on Worrall's attempt in his initial presentation of structural realism to cash out the notion of "structure" by reference to mathematical equations specifically. These problems are avoided, however, by more recent presentations of structural realism which cash out "structure" in terms of relations between entities more generally. This is a major conclusion of Newman's (2004) paper, in which he goes on to endorse the "Ramsey-sentence realism" approach due to Cruse and Papineau (2002; Cruse, 2005; Papineau, 2010). Newman was apparently unaware, however, that Worrall (2007; Worrall & Zahar, 2001) himself had, by this time, *also* explicitly endorsed a Ramsey-sentence approach. In any case, this is the formulation of structural realism that will be discussed for the remainder of this section.

Ramsey (1931) gives the recipe for formulating the Ramsey-sentence of a theory. In very general terms, a scientific theory attributes certain properties to a collection of observable and unobservable (theoretical) entities, and asserts that certain relations hold between these entities. We can therefore express a given theory as: $T(t_1, t_2, \ldots t_n; o_1, o_2, \ldots o_m)$, where T is a complex predicate expressing all the first-order properties and relations, the $t_i$ are theoretical terms, and the $o_i$ are observational terms. To avoid referring to named theoretical entities, we now substitute all the theoretical terms in this expression with variables, and quantify over the variables to form the Ramsey-sentence: $\exists u_1 \exists u_{2\ldots} \exists u_n T(u_1, u_2, \ldots u_n; o_1, o_2, \ldots o_m)$. This sentence claims that *there are* certain unobservable entities which are related to each other and to named observable phenomena in a particular way, but is not committed on exactly what these entities are.

Notice two things about this formulation of structural realism. Firstly, it is in perfect agreement with Psillos' view on the relationship between "structure" and "nature" – the nature of a theoretical entity is specified *exactly* by the role it plays in the theory! Secondly, it is committed to an extreme descriptive account of reference, and so stands in direct opposition to the causal accounts discussed in section 2 above. This commitment is explicitly defended by (Worrall, 2007, pp. 148-149). He argues that attempting to apply a causal theory of reference requires the nonsensical presupposition that we can "stand outside" our theories and directly observe how they match up with the world. Structural realism, in contrast, takes the view that our knowledge of the world is necessarily mediated by theory, and thus by description.

With the "structure" of a theory cashed out more rigorously in terms of set-theoretic relations, the notion of structural continuity between two theories can now be formalised along the same lines. In set theory, a structure S is defined as an object consisting of a set of entities X and a set of relations between those entities Q. We thus write S = {X, Q}. A structure S is isomorphic to another structure T = {Y, R} just in case there is a one-to-one function *f* that maps every entity in the set X to an entity in Y in such a way that, for any ordered tuple of entities in X, call it ($x_1$, $x_2$, ... $x_n$), satisfying the relation Q, there is a corresponding ordered tuple in Y, ($f(x_1)$, $f(x_2)$, ... $f(x_n)$), satisfying the relation R, and vice versa. For instance, the set of natural numbers is isomorphic to the set of even natural numbers because we can define a one-to-one mapping (e.g. $f(x) = 2x$) so that a given relation that holds in the former (e.g. the greater-than ordering) has a corresponding relation that holds in the latter (also the greater-than ordering). To be a structural realist, then, is to claim that there exists an isomorphism between the set-theoretic characterisation of a successful scientific theory and the mind-independent world, and consequently that any future successful theory in this particular empirical domain will also be isomorphic to this theory.

I shall now address two major objections to this view. The first is generally referred to as the “Newman objection”, since it originates in Newman’s criticism of Russell’s (Russell, 1992)/1927) *Analysis of Matter*. The argument has been revived more recently by Demopoulos and Friedman (1985) in response to Maxwell’s revival of Russell’s approach. The argument and the responses to it have been very ably summarised recently by Ainsworth (2009) and I largely follow his treatment here. At the heart of Newman’s paper is the claim that:

> “Any collection of things can be organised so as to have the structure W [where W is any chosen structure], provided there are the right number of them. Hence the doctrine that *only* structure is known involves the doctrine that *nothing* can be known that is not logically deducible from the mere fact of existence, except (‘theoretically’) the number of constituting objects.” (Newman, 1928, p. 144, emphasis in original).

In other words, inferring that the world must belong to a particular isomorphism class doesn’t tell us very much, since on any set with enough objects in it, we can define a set of relations so that the resulting structure satisfies the isomorphism.

Some care should be taken in how this result is interpreted. Recall that a Ramsey-sentence makes assertions involving both theoretical and observable terms. If the Newman objection is well-taken, the assertions involving theoretical terms are satisfied provided there are at least as many entities in the world as there are distinct theoretical terms. So to accept a Ramsey-sentence is just to accept the claims of a theory about observables, plus an effectively trivial claim about the number of things there are in the world. Structural realism is therefore in fact equivalent to empiricist anti-realism, along the lines of van Fraassen’s constructive empiricism (see Ketland, 2004 for a more formal and detailed version of this argument).

Several responses have been offered to this argument, but I shall focus on those due to Worrall and Zahar (Worrall, 2007; Worrall & Zahar, 2001; Zahar, 2004). One response is that, even if assertions expressed purely in the theoretical vocabulary are rendered trivial, sentences in the observational vocabulary are nevertheless extremely expressive:

> “Many sentences expressed in purely observational vocabulary should count, in any one’s book, as theoretical: one often cited example is the claim that there are unobservables – that is, individuals possessing no observable property.” (Worrall, 2007, p. 152)

This response, however, surely misses the point. The question is whether the structural realist is able to distinguish her position from that of the constructive empiricist. And the latter will, I think, agree that claims which are not verifiable by direct observation but which are expressed in observational vocabulary are proper objects of scientific belief.

A more convincing response is as follows:

> “...if we follow Quine’s dictum that ‘to be is to be quantified over’ (or, better: ‘to be asserted to be is to be quantified over in a sentence that you assert’) then the Ramsey sentence of some theory T ($S_1$,... $S_n$, $O_1$,... $O_r$) clearly asserts that the ‘natural kinds’ $S_1$, ..., $S_n$ (the extensions of the theoretical predicates S1, ... S*n* in the initial theory T) exist in reality just as realists want to say. It is just that – as always – we fool ourselves if we think that we have any independent grip on what the $S_1$, ..., $S_n$ are aside from whatever it is that satisfy the Ramsey sentence (assuming that the theory whose Ramsey sentence we are considering is true).” (*ibid.*, p. 152)

The Ramsey-sentence, then, asserts more than just that there are a certain number of entities exist in the world, and that there are certain regularities expressible in observational vocabulary. It asserts that certain entities exist, and that they act to structure the phenomenal world in a certain way. This suggestion, as it stands, is rather rudimentary and should be developed in more detail. But it is nevertheless clear how this formulation at least potentially satisfies the demand for a prospective account of essentialness. To say that there exists an entity which is related to various observable phenomena, and other theoretical entities, in a particular way is to imply that any empirically successful future theory will posit an entity which satisfies the same relations.

There is, however, a relatively decisive criticism of structural realism. This is suggested by Chakravartty:

> “... current theories do not retain all of the structures described by their predecessors. But then not all structures are causally connected in appropriate ways to our practices of detection.” (Chakravartty, 2004, p. 164)

and Stanford:

> “Francis Galton's ancestral law of inheritance, for instance, was the central mathematical formalism and the most predictively successful aspect (see below) of his "stirp" theory of inheritance... By present lights, it would be extremely misleading, if not outright mistaken, to say that even the mathematical structure expressed by Galton's Ancestral Law is preserved in contemporary genetics.” (Stanford, 2003a, pp. 570-571)

So, if we are concerned with which theoretical elements are essential to the predictive success of the theory and preserved in successor theories, even the narrow equation-focussed version of structural realism is too inclusive. And this problem becomes even worse when we adopt the even more inclusive Ramsey-sentence criterion.

Returning to the Fresnel example, notice that when we start cashing out structure in terms of relations between entities generally there is no difference *in kind* between Fresnel’s equations and the posit of a luminiferous ether. All the theoretical entities are asserted to participate in relations, and some of these relations include observables. In particular, the existence of a luminiferous ether means that an observer will measure a different speed of light depending on which direction she is travelling relative to it (the “ether wind” effect). The posit of a luminiferous ether, in other words, gives rise to additional empirical consequences, and these also follow from the Ramsey-sentence of a theory that includes this posit.

The reasoning of structural realists appears to have been as follows. First, it is asserted that the distinction between mathematical formalism and other types of theoretical posit is co-extensive with that that between essential and inessential posits. Secondly, the emphasis of structural realism is now shifted from mathematical equations to relations more generally, under the assumption that the same distinctions between essential and inessential would be preserved. This reasoning can be challenged in both stages. The initial assertion is challenged by Stanford’s example given above. And the assumption underpinning the second stage is undermined by various additional counterexamples. The point of all this is not to claim that the luminiferous ether *should* be counted as essential. Rather, it is to point out that the most interesting difference between Fresnel’s equations and the luminiferous ether is surely not the mere fact that one is expressed in equations and the other is not. What *is* interesting is that, while the empirical consequences which follow from Fresnel’s equations are amply borne out by experiment, the empirical consequences attributable to the luminiferous ether simply fail to emerge (as in the famous Michelson-Morley experiment; discussed by Swenson, 1972).

So, the form of structural realism presented thus far attempts to pick out what is essential to a theory’s predictive successes by isolating those theoretical posits which assert some structural relation to hold between observables. But it is silent on the seemingly obvious point that we must also determine which relations are actually *confirmed* to hold by some empirical evidence. When stated so baldly this approach is obviously perverse, and it is doubtful that this is what Worrall *et al.*

actually intend. We surely need to know just which of the empirical consequences of the theory *are* the empirical successes before we can determine which parts of the theory are required to achieve them. I therefore turn to versions of partial realism which pay more attention to how particular theoretical posits are related to particular empirical consequences.

## 6. Semi-realism

Chakravartty (Chakravartty, 1998; 2003; 2004; 2007) has argued that entity realism and structural realism mutually imply one another and so are equivalent. He suggests the term "semi-realism" for the overall view. In this section, I will briefly examine Chakravartty's argument that the two views in question are equivalent, and investigate whether his arguments are able to overcome the problems with these views outlined above.

Chakravartty gives a version of the argument, seen in section 4 above, that we cannot sensibly talk about picking out certain entities without at least describing some of their properties. His reaction to this argument, which parallels Cartwright's distinction between causal and theoretical explanation, is as follows:

> "We infer entity existence on the basis of perceptions ground upon certain causal regularities having to do with interactions between objects. Let us define *detection properties* as those upon which the causal regularities of our detections depend, or in virtue of which those regularities are manifested. *Auxiliary properties*, then, are those associated with the object under consideration, but not essential (in the sense that we do not appeal to them) in establishing existence claims... Theories enumerate both detection and auxiliary properties of entities, but only the former are tied to perceptual experience." (Chakravartty, 1998, pp. 394-395, emphasis in original)

To pick out the detection properties of some posited theoretical entity, we must ask which properties are required to be instantiated in order for the observed phenomenal regularities to occur. Chakravartty, like many other partial realists, illustrates this reasoning by reference to Fresnel's equations. He argues that these equations "... demand some kind of influence, propagated rectilinearly and resolvable into two components, oscillating at right angles to one another and to the direction of propagation. The property or properties of light in virtue of which such influences are realized are detection properties." (*ibid.*, p. 396)

Chakravartty also anticipates Worrall's point, made in section 5, that structural realism is, by Quinean ontological reasoning, committed to the truth of existential claims about certain unobservable entities. So far, then, semi-realism seems just equivalent to the structural realism – it is committed to the existence of certain unobservable entities which act to produce certain well-confirmed regularities in the observable world. Chakravartty points out, however, that insofar as structural realism is committed to the existence of unobservable entities, it entails a sort of entity realism. It is unclear whether it entails *Hacking's* entity realism, since the latter is also committed to the causal theory of reference, and Worrall's version of structural realism is explicitly committed to a descriptivist theory of reference. But let us suppose that we are in any case presently concerned with a more descriptivist version of entity realism, since we have in any case acknowledged the need for the attribution of at least some properties in order to characterise an entity.

Perhaps it is not surprising that structural realism entails entity realism, since the former is in some sense a more 'ambitious' form of realism. A relation of entailment in the other direction, however, would seem to be more contentious. Chakravartty nevertheless argues as follows. Belief in entities under entity realism is underpinned by the postulation of certain relations ("detection properties"). Moreover, we would expect these relations still to be found in subsequent theories which attempt to describe the same entities. So, because entity realism is committed to the existence of certain relations which are preserved across theory change, it is committed to structural realism.

Chakravartty himself suggests the most obvious objection to this claim, and gives a response to it:

> “It may be objection that surely not *all* structures have to with causal relations involving the detection properties of entities. Clearly we can imagine different kinds of structures, such as ones linking auxiliary properties. But for the advocate of [structural realism], such flights of fancy are not particularly helpful, for not just any structure will do. [Structural realism] requires *stable* structures – ones which are, in fact, likely to be preserved.” (Chakravartty, 1998, p. 400, emphasis in original)

This clearly mirrors the complaint, in section 5 above, that the structural realist would seem to be committed to all kinds of ‘superfluous’ structure. However, whereas I argue that the structural realist requires some *additional* criterion by which to pick out only the ‘essential’ structure, Chakravartty appears to believe that the structural realist is *already* committed to the essentialness of “detection properties”.

The major problem with Chakravartty’s proposed identification of structural realism with entity realism is that it appears to trade on an equivocation in the notion of a “detection property”. At least in the first instance, Chakravartty wishes to identify detection properties with the “low-level causal properties” of entities which Hacking argues we appeal to when we successfully manipulate these entities. Hacking does not, however, provide a detailed account of which of the theoretical posits in actual scientific theories are to be understood as describing these “low-level causal properties”. It is nevertheless reasonable to assume that the distinction between such low-level properties and more abstract theoretical properties maps onto Cartwright’s distinction between phenomenological and fundamental laws. But if detection properties are those that underlie descriptively accurate phenomenological laws, then the structural realist is a realist about more than just detection properties. Chakravartty’s use of the structural realist’s favourite illustrative example, namely Fresnel’s equations, is therefore puzzling. These equations do not directly describe any actual observable state of affairs; they are in fact abstract rules which are used (often very successfully) to derive phenomenological models for particular cases. So “detection property” can be given either a limited or an expansive reading, but the distinction between low-level phenomenological laws and abstract theoretical posits cannot simply be elided. Chakravartty’s semi-realism, therefore, can be either a form of entity/phenomenological realism or a form of structural realism, but not both.

## **7. Working posits**

Both the “working posits” idea due to Kitcher (1993) and the “divide et impera” strategy due to Psillos (1999) claim, in broad terms, that the essential posits of a successful theory are those that are involved in the actual derivations of successful predictions. I will outline each of these views in term, before examining some criticisms of the broad position.

Kitcher (1993) presents the concept of “problem-solving schemata”. These are generic sentences, each of which expresses some regular pattern of reasoning found in a particular scientific theory. The primary ingredient of a schema is a sentence containing “dummy letters” at certain locations. The “filling instructions” of the schema tell us what sort of entities can be put in place of the dummies to create a potential explanation. Consider, for instance, the schematic sentence "Organisms homozygous for *A* develop P" found in classical genetics. The filling instructions say that the name of an allele should be put in place of *A*, whereas *P* should be replaced by the name of a phenotypic trait (ibid, p. 82). Kitcher views these schemata as the major workhorses of science, producing explanations and predictions for particular observed occurrences of phenomena.

Using this concept, Kitcher goes on to state the working posits idea as follows:

> “Distinguish two kinds of posits introduced within scientific practice, working posits (the putative referents of terms that occur in problem-solving

> schemata) and presuppositional posits (those entities that apparently have to exist if the instances of the schemata are to be true).... The moral of Laudan's story [in the 1981 paper] is not that theoretical positing in general is untrustworthy, but that presuppositional posits are suspect. The ether is a prime example of a presuppositional posit, rarely employed in explanation or prediction, never subjected to empirical measurement ..., yet seemingly required to exist if the claims about electromagnetic and light waves were to be true." (ibid, p. 149)

The essential elements of a theory, therefore, are none other than those actually found in problem-solving schemata[12]. Notice that Kitcher also uses the Fresnel case study, but gives a different rationale to Worrall for regarding the luminiferous ether as inessential.

Before examining Psillos' positive account, it is interesting to note that he sets the stage by attacking Kitcher's account, and does so by anticipating Stanford's complaint against retrospective variants of partial realism. Psillos focuses on a comment made by Kitcher:

> "But for Fresnel and many of those who followed him, the existence of such an ether was a presupposition of the successful schemata for treating interference, diffraction, and polarization, apparently forced upon wave theorists by their belief that any wave propagation requires a medium in which the wave propagates. All the successes of the schema can be preserved, even if the belief and the presupposition that it brings in its train are abandoned." (Kitcher, 1993, p. 145)

From this passage, Psillos interprets the major distinction between working and presuppositional posits to be that only the latter are "eliminable without derivational loss". He then argues that "[T]his suggestion is retroactive and open to the charge

[12] Though recall from the discussion in section 3 that Kitcher's account also partly depends on a causal theory of reference. In this section, the focused is narrowed to the more "descriptive" part of Kitcher and Psillos' accounts.

that it is ad hoc: the eliminable posits are those that get abandoned" (Psillos, 1999, p. 106). Incidentally, Psillos himself in certain passages can be read as offering either retrospective or prospective accounts. However, the fact that he criticises Kitcher on precisely the grounds that he interprets the latter as endorsing a retrospective account certainly suggests that he does not in fact endorse such an account himself.

It is, however, also a mistake on Psillos' part to read Kitcher as endorsing a retrospective account. Kitcher does indeed state that paradigmatic presuppositional posits, such as Fresnel's ether, tend to be progressively eliminated in later rounds of theorising without any loss of empirical power. However, this is not the primary means by which he picks out inessential elements. Given the importance he attaches to explanatory schemata in earlier parts of this book, it is much more natural to interpret the definition of working and presuppositional posits in terms of these schemata as the primary one. Eliminability without derivational loss therefore emerges as a *consequence* of the posits in question being inessential; it is not *definitive* of inessentiality. In any case, as I argue in the following section, the criterion of eliminability without derivational loss is not necessarily applicable only with hindsight.

His misinterpretation of Kitcher notwithstanding, Psillos articulates his *divide et impera* strategy as follows:

> "My claim is that it is precisely those theoretical constituents which scientists themselves believed to contribute to the successes of their theories (and hence to be supported by the evidence) that tend to get retained in theory change. Whereas, the constituents that do not 'carry-over' tend to be those that scientists themselves considered too speculative and unsupported to be taken seriously." (Psillos, 1999, p. 107)

So both Kitcher and Psillos claim that the essential elements of a theory are those *actually used* by scientists to derive results. Despite some remaining differences between the two approaches, I will therefore address my remarks in the remainder of the section to this underlying idea.

A major problem with this view is that it is still extremely unclear how we are to distinguish working and idle posits. In the remainder of this and the following section, I will offer several plausible interpretations of “working posit”, and ultimately argue that none is satisfactory, thus setting the stage for my own view in section 10. It is possible to start getting a sense of the difficulties of interpretation by way of a particularly illuminating exchange between Psillos and Chang, centred on the science of heat in the early nineteenth century. The dominant research programme at this time posited that heat was constituted by a material substance called “caloric”. This programme achieved several indisputably successful explanations and predictions of empirical phenomena, notably Laplace’s remarkably accurate derivation of the speed of sound in air. And yet we now believe that caloric does not exist, so this case represents an instance of the classic PMI-type challenge.

Psillos (1999, ch. 5) attempts to show that the existence of caloric was not a working posit for Laplace by focusing on the details of his derivation. This begins by postulating that a region of air experiences isothermal compression when the pressure of the sound wave initially strikes it. This is followed immediately by adiabatic heating as the compression causes the temperature of this region to increase[13]. Psillos notes that this derivation does not explicitly appeal to any microphysical properties of the system, and indeed that the macrophysical “explanation of the propagation of sound in terms of an adiabatic process is essentially correct and has been retained in the subsequent theoretical accounts of heat” (*ibid.*, p. 115). Hence positing the existence of a substance with the particular microphysical properties attributed to caloric is not essential to the explanation.

[13] “Isothermal” denotes any process occurring at constant temperature. “Adiabatic” denotes a process that involves no transfer of heat between the system and its surroundings. In Laplace’s derivation, heating is assumed to be adiabatic because it is so rapid that there is no opportunity for heat to be transferred.

Chang (2003) responds by arguing that Laplace's mechanistic reasoning about these macrophysical processes was crucially underpinned by "assumptions about the material nature of caloric". Laplace thought of caloric as a "subtle fluid", the particles of which are attracted to ordinary matter but experience mutual repulsion as a function of proximity. Temperature is a measure of the amount of "free" caloric in a region of space (i.e. not bound to ordinary matter). This accounts for the temperature increase during adiabatic compression of a gas, as the mutual repulsion between bound caloric particles forces them into the empty space surrounding the gas molecules.

Chang implicitly understands "essential" elements as those that are *causally* essential for the derivation. Since he (quite plausibly) doubts that Laplace could have conceived of the necessary macroscopic processes had he not believed some story about the underlying microphysics, he therefore regards the microphysical posits as essential. Psillos differs from Chang's causal interpretation in that he focuses on the theoretical posits that Laplace was sufficiently confident about that he was willing to explicitly invoke them when deriving his result. So, both authors appeal to some notion of "working posits" to determine which elements are essential for the derivation of an empirical result. Speaking very roughly, we can say that Psillos interprets this notion *logically* or even *psychologically*, whereas Chang interprets it *causally*. There are, in fact, several other possible interpretations of the working posits idea. I will address these, and critically assess their adequacy, in the following section.

## **8. Differing interpretations of the working posits idea**

The two competing interpretations of the working posits idea due to Chang and Psillos do not remotely exhaust the space of possible interpretations[14]. In this and the following section, this space is explored in more detail. The first interpretation

[14] A large part of the following list was filled in with the help of participants at the Physics and Philosophy of Kirchhoff's Theory of Diffraction colloquium held at Durham University on the 29th of May 2012, particularly Peter Vickers and Juha Saatsi.

to consider is that attributed to Chang in the discussion above, namely that (i) a working posit is that which causally contributes to a successful derivation. The second interpretation is that attributed by Psillos to Kitcher, namely that (ii) a working posit is that which is not eliminable without derivational loss. The third interpretation has not been discussed in the literature, but was suggested by Pete Vickers in conversation. It states that (iii) the working posits in a derivation are those which cannot themselves be derived without appealing to more than one logically prior assumption, and any posits logically derivable from these. The fourth and fifth interpretations are both found in Psillos, and are closely related to each other. They are nevertheless worth distinguishing clearly. The fourth interpretation states that (iv) a working posit is that which the scientists performing a derivation believe to contribute to the success of this derivation. The fifth interpretation states that (v) a working posit is that which is explicitly invoked in the formal derivation of a successful result. The sixth interpretation (vi) is a hybrid of some of these other positions, and has recently advocated by Vickers (forthcoming). This sixth interpretation is addressed by itself in section 9.

Before examining each of these interpretations it detail, it is worthwhile recalling the motivation behind this exercise. The interpretation that emerges from this discussion is supposed to pick out those posits of a theory which, if true, will *explain* the success of the theory or derivation in which they feature. Assuming the soundness of the NMA, belief in the truth of these posits would then be justified. So, because the following assessment of the various interpretations relies on the notion of explanation, it is worth making a few remarks about it before proceeding. The first point to be made is that there are at least two kinds of explanation. The explanandum of a c*ausal* explanation is an event, and the explanation attempts to outline some sequence of prior events that results, by physical processes, in the occurrence of the explanandum. A car accident might, for instance, be explained by noting, among other things, that the car was travelling fast and that its brakes were faulty. The explanandum of a *logical* explanation is the assertion of a proposition by some people, and the explanation attempts to trace some process of reasoning undertaken by these people of which the explanandum is the justified conclusion. Jane’s belief that all ravens are black, for instance, is explained by her observation of several black ravens and her allegiance to the principle of induction.

The second point to be made is that a successful explanation, either causal or logical, generally makes reference (often implicitly) to a “foil” or possible scenario in which the explanandum did not occur, or occurred in a different way (Lipton, 1990). For instance, one cannot simply explain why the leaves on the trees turn yellow in November; any sensible explanation proffered will actually explain, say, why they turn yellow in November *rather* than in January. Similarly, a satisfactory explanation of why Jane believes that all ravens are black will explain why she has this belief as opposed to the belief that they are all white, or to not having any belief on this question at all.

The advantage of the causal interpretation (i) is that we generally have at least an intuitive grasp of causal explanation. So, if it is the case that certain theoretical posits are causally essential to the success of a derivation, it is relatively straightforward to explain how these posits are responsible for that success. The problem, from the realist perspective, is that this explanation need not appeal in any way to the *truth* of these posits. That they are part of the causal chain leading to the success is sufficient. Thus Fresnel’s belief in a luminiferous ether and Laplace’s belief in caloric are both counted as essential to the success of their respective theories. This, of course, is grist to the anti-realist’s mill.

The causal interpretation, however, casts its net even more widely than this type of contentious metaphysical posit. For instance, since neither Fresnel nor Laplace could have theorised successfully without being able to breathe, a causal interpretation would also seem to count as essential the presence of oxygen in the atmosphere surrounding them while they were working. Of course, even the anti-realist is primarily concerned with posits that have actual propositional content. But, even supposing the discussion is limited to theoretical beliefs, under the causal account:

> “Credit will have to be attributed to all [causally] responsible constituents, including mere heuristics (such as mystical beliefs), weak analogies, mistaken calculations, logically invalid reasoning, etc.” (Lyons, 2006, p. 543)

Indeed, it is very often the case that posits which scientists *themselves* quickly realise are mistaken nevertheless direct those scientists to fertile ground. The point is that, even before considering examples like Chang’s, no realist who has given the matter any thought would sign on to the causal account in the first place. Causal responsibility therefore cannot be the right way to interpret the notion of a “working posit”. An adequate interpretation of this notion must instead invoke some sort of logical explanation.

The simplest purely logical interpretation (ii) states that essential elements are just those which cannot be is eliminated without derivational loss. As alluded to in the previous section, Psillos’ claim that this interpretation can only be applied retrospectively is unsound. Although certain logical relations will only become apparent with time, it is nevertheless possible *at a purely logical level* to identify which assumptions are required for a given derivation to proceed. To give an example, the proposition that light propagates in the form of a transverse wave follows logically from the assumption that an elastic solid luminiferous ether exists. It is nevertheless the case that the same results can be derived if the ether posit is eliminated and it is simply assumed that light propagates as a transverse wave. The ether posit is thus eliminable without derivational loss and so is not a working posit.

The problems with this account become evident upon closer examination of the actual counterfactual scenario invoked when one is asked to judge that some posit is eliminable without derivational loss. Specifically, it seems that this scenario not only eliminates the posit in question but assumes at least one of its logical consequences. However, if it is legitimate to take an intermediate result as an assumption and regard as inessential the assumptions used to derive it, the inevitable consequence is that almost nothing *is* essential. This process of elimination can be continued until we take the final result itself as an assumption and regard all other theoretical posits leading up to it as inessential. Everything except the most concrete statement of the phenomenon to be explained, in fact, can be eliminated without sacrificing the “derivation” of this statement. This is a *reductio* of this interpretation of the working posits approach, which must consequently be abandoned[15].

Since the third interpretation (iii) is relatively abstract and has not yet been proposed in the literature, a simple analogy will give some sense of what it requires. Imagine a theoretical derivation as a river of propositions, with the initial assumptions as tributaries. The initial assumptions ‘converge’ at ‘confluence points’ to yield intermediate results, and the entire system eventually converges at the ‘main stem’, representing the final result. Under this interpretation, the essential posits of a derivation are those propositions on which several logically prior assumptions converge and the propositions which follow from these. Inessential posits, in contrast, are those which are derivable from a single logically prior assumption. Continuing with the river metaphor, the idea is that the logical branching structure of the derivation is essential, but the tributaries can be pared away until the outermost confluence points.

Vickers proposed this interpretation as a way to avoid the catastrophic outcome of adopting the simple logical interpretation (ii). There is also at least some intuitive reason for regarding a posit that is a ‘confluence point’ as having greater explanatory weight than one which (in the context of the particular derivation) is merely a ‘tributary’. The intuition is that the real work of the derivation involves the combination of assumptions to give novel theoretical results. We can say of an assumption that is logically ‘upstream’ of a confluence point something to the effect that its ‘only function’ is to give rise to other assumptions which are actually used in the derivation.

[15] Though notice that, depending on how the “final result” of the derivation is defined, this endpoint may in fact be very similar to phenomenological realism.

Although this interpretation is intuitively appealing, it is at least incomplete. It may certainly be true within the scope of a *single* derivation that a given posit can be eliminated without altering the branching structure of the main argument. A given theoretical posit may, however, be involved in several derivations. The caloric posit is, it seems, eliminable from Laplace’s derivation of the speed of sound along the lines suggested by this interpretation. But this posit was also involved in other derivations in the broader theory of heat. From the perspective of the single derivation, then, the posit is an isolated tributary, the truth of which explains very little. But from the perspective of the broader theory, its truth explains a great deal. This point is addressed in more detail in section 10, where a positive account of essentialness is proposed.

The fourth interpretation (iv) involves a relatively straightforward appeal to the judgement of working scientists. However, it should be obvious immediately that there is a historiographical problem inherent in applying this interpretation. Scientists will not always publicly distinguish which elements of their theories they are more or less confident about. So frequently one will have to *infer* that a scientist is more confident about a given element from the fact that she uses it in derivations in a certain way. This is likely why Psillos frequently slips between the purely psychological interpretation (iv) that he offers initially and a more logical or historical interpretation (v) that depends on which posits scientists explicitly invoke.

Putting such practical issues aside, it is still worth underlining that a purely psychological interpretation cannot be adequate for the realist. Firstly, there are simply too many obvious counterexamples. It is true that scientists will often hedge their commitments to more abstract theoretical posits, especially if there are other extant theories in competition with their preferred account. As Psillos himself points out, this was exactly the case with Laplace, who was aware of the “vibratory” theory of heat, even if he rejected it. There are also, however, cases where scientists very confidently endorsed abstract posits which turned out, from the perspective of later theories, to be false. Fresnel and other early nineteenth-century physicists’ attitude towards the luminiferous ether is in fact one such case, especially since at that time it was difficult even to *conceive* of the propagation of a wave without the presence of a mechanical medium (see the more detailed discussion of this case in Chapter 2.3 and section 5 of this chapter).

A second, related, argument against the psychological interpretation is that it is, so to speak, intellectually incurious. Stating that scientists were committed to certain theoretical posits and that from these certain empirical results can be inferred does, indeed, explain the empirical success of the theory. But stated thusly, this is no different from the simple causal interpretation (i). The psychological interpretation draws its force from the notion that scientists have some genuine insight into what parts of their theories are “actually responsible” for these empirical successes. So, supposing this is true, a complete explanation of the empirical success of a theory should address the criteria scientists apply when they make such judgements, and demonstrate that these are *appropriate* criteria. The psychological interpretation must therefore be supplemented with some more substantial logical interpretation.

The fifth interpretation (v) focuses on those posits which are explicitly invoked in the derivation of a result. This does make some *prima facie* sense. If we are concerned with explaining how the successful result was obtained, it seems sensible to focus on the actual reasoning of the scientists involved, independently of what other beliefs they may have had ‘in the background’. Notice, however, that if we include posits which are explicitly invoked at any stage in a scientist’s reasoning, then we are again stuck with the elasticity of the ether and so on. One way around this is to draw a distinction between reasoning in general and the *derivation* of a result, considered relatively narrowly[16].

But then this interpretation seems to be committed to the notion that scientific reasoning comes in two “modes”. Scientists start off doing some informal, intuitive reasoning to motivate assumptions, and this reasoning will often involve more abstract and/or metaphysical assumptions. Then they finish with the informal

[16] This distinction occurred to me in the course of a conversation with Juha Saatsi, and I cannot recall to which of us the credit belongs.

reasoning, write down some assumptions and get going with a formal derivation. The contention that this distinction is explanatorily relevant is, however, highly problematic. The first problem is that the boundary between the two modes of reasoning is deeply contingent. A major factor determining when scientists will apply formal reasoning is the complexity of the problem in question. Problems that involve too many working parts to be stored in a human being's working memory need to be simplified *via* the use of symbols, written down, and broken into collections of smaller problems that can be worked through piecemeal. But where informal reasoning stops and formal reasoning takes over is therefore contingent on the reasoning capacities of the person or people involved. Someone with a well-developed 'physical intuition' will, in constructing a model for a particular circumstance, be able to write down fairly complex assumptions straightaway. In contrast, a lesser scientist may need to start with more basic assumptions and derive the more accomplished scientist's initial assumptions as intermediate results. The exact point in a chain of scientific reasoning where the formal "derivation" begins is therefore too contingent on the particulars of the scientists involved to bear the explanatory weight that this interpretation would place on it.

The second problem is that, even assuming that this distinction between modes of reasoning can be made, it is unclear why it is explanatorily relevant. If one is concerned with a logical explanation for some belief, it seems that one should be interested in how the believer reasoned, irrespective of whether this reasoning was formal or informal. As argued in Chapter 2.10, all else being equal, the truth of theoretical posit is certainly a *better* explanation of the success of an empirical prediction if there is a deductive relationship between the two. But this does not imply that, if the relationship is weaker than deductive entailment, the truth of the posit *cannot* explain the success of the theory sufficiently well that the explanation is an instance of the NMA.

## 9. Essentially contributing parts of derivation internal posits

The sixth interpretation (vi) has recently been advanced by Vickers (Vickers, forthcoming), and represents the most sophisticated attempt yet to cash out the notion of a 'working posit'. I will thus devote considerably more analytical attention to it than I have to the other interpretations outlined above. Vickers' interpretation can be characterised as an attempt to identify a *logically minimal* set of posits that are *actually involved* in the derivation of a successful result, and as such represents a synthesis of interpretations (ii) and (v) given above. Vickers begins by attempting to distinguish more clearly between "derivation internal posits" (DIPs) and "derivation external posits" (DEPs). Unlike Psillos, he does not draw this distinction on the basis of which posits scientists invoke explicitly, but on the type of logical connection the respective posits have to the final result. For Vickers, the DIPs are those which (possibly together with other posits) eventually result in the final result *via* "truth-preserving inference", whereas DEPs at best "influence" or "inspire" scientists to adopt other posits.

His illustrates the distinction between DIPs and DEPs by discussing the construction of the first "achromatic lenses" for telescopes. A normal convex lens will differentially refract different wavelengths of incoming light and so produce "chromatic aberrations" in the resulting image. An achromatic lens is constructed by pairing a convex and a concave lens of different materials in such a way that the overall lens is still convex (and so focuses light), but does not differentially refract different wavelengths of light. Interestingly, this design was directly inspired by the structure of the human eye, in which light passes through several distinct materials (the lens and the humours) before being captured at the retina. However, as it turns out, and contrary to the belief of the scientists concerned, the human eye does *not* eliminate chromatic aberration. Instead, the brain simply compensates for this effect. Nevertheless, Vickers argues, although the assumption that the eye eliminates chromatic aberration turned out to be false, we should not in any case consider this a working posit in the derivation of the accurate empirical prediction. Insofar as this assumption contributes to this prediction, it is only by means of the vague suggestion that a construction of the appropriate type is possible. It is thus a DEP. The detailed, truth-preserving calculations begin only with more specific assumptions about the refractive properties of the materials involved in the actual construction of the telescope lens. These assumptions are therefore the DIPs.

Vicker's further distinguishes the "essentially contributing parts" (ECPs) of the DIPs, which he identifies as the true working posits of the theory. To find the ECPs of a given DIP, we are enjoined to ask whether the derivation in which this posit is involved could have proceeded by assuming, instead of the actual posit used, some weaker posit which is a consequence of it. To illustrate this point, Vickers returns to a paper he co-authored with Saatsi concerning Kirchhoff's diffraction formula (Saatsi & Vickers, 2010). This formula produces accurate predictions for the diffraction pattern of light which has passed through an aperture in an opaque screen. Saatsi and Vickers' concern is that one of the crucial assumptions, a DIP in fact, in the derivation of the formula is that the intensity of the light in the aperture is as if there was no screen at all. This assumption is manifestly false from the perspective of modern optics. According to Vickers, however, we can identify an ECP of this DIP which is not compromised from our modern perspective. Drawing on work by Brooker (2008), he notes that, the intensity of light at the aperture under Kirchhoff's assumption can be represented by a Fourier function with infinitely many terms. As it happens, many of these terms make only negligible contributions to the final predicted diffraction pattern. Taking only the non-negligible terms, it becomes apparent that they together give a picture of the behaviour of light at the aperture which is quantitatively nearly identical to that which is predicted by our modern theory of optics. So Kirchhoff's assumption, although not a working posit itself, has as a consequence a weaker claim which *is* a working posit.

As emphasised above, Vickers interpretation of the working posits idea is the most sophisticated yet to appear in the scientific realism literature, and is extremely appealing. Ultimately, however, it is not satisfactory. Consider first Vickers' idea of identifying the working posits *within* DIPs by picking out the ECPs of these posits. An obvious objection to this is the same as that applied to interpretation (ii). If one is simply allowed to take logical *consequences* of particular posits to be the ECPs of those posits, then what is to stop one from simply regarding all but the most low-level empirical claims of a theory as “inessential”? The only apparent solution to this problem relies on the major difference between the present proposition and interpretation (ii), namely that in Vickers interpretation the DIPs are already specified. The challenge is only to pick out the logically minimal parts of a given set of posits, not of the overall derivation. But this response just leads to the question of which posits, exactly, are the DIPs, which is where the true difficulties with Vickers’ account lie.

Vickers is more convincing than interpretation (v) in attempting to distinguish two distinct ‘modes’ of scientific reasoning. Nevertheless, the earlier objection – that the line where informal reasoning ends and formal reasoning begins is indistinct and contingent – still has bite. A well-trained scientist, for instance, might through intuition be able to skip various steps in a derivation that a lesser practitioner would have to go through explicitly and formally. It seems that Vickers is getting at a similar point when he remarks:

> “Suppose (as here) we are only considering deductive derivations. Nevertheless, we might want to label some posit *internal* on the grounds that there are implicit posits which logically connect it to the other derivation-internal posits. In other words, what on the surface looks like mere *influence* can be turned into an *inferential* relationship if other ‘implicit’ posits are added.” (Vickers, forthcoming, emphasis in original)

The distinction between DIPs and DEPs is, in other words, not as clear-cut as initially suggested.

Although, Vickers largely puts this line of criticism to one side in the forthcoming paper, he has addressed it briefly elsewhere:

> “My claim now, in light of these considerations, is simply this. So long as it is clear-cut that a given posit is ‘internal’, then the realist should consider making a commitment to ECPs etc. The line between internal and external may well be vague, and the realist can just (for now) stay silent about the vague area. The challenge I am taking up first and foremost is to try to express how the realist should commit in cases where posits in a derivation clearly *are* connected to the prediction in a deductive chain. We go upstream so far, and stay silent about a possible realist commitment to posits any further upstream than that.” (Vickers, 2012, personal communication)

So, although the line between DIPs and DEPs is “vague”, there are nevertheless many cases where a given posit is clearly connected to an empirical result by a chain of deductive reasoning. So a realist commitment under his account is only justified in respect of “at least *some* posits in at least *some* cases” (*ibid.*, his emphasis). This is highly reminiscent of van Fraassen’s argument in response to the criticism that “observable” is a vague term:

> “In Sextus Empiricus, we find the argument that incest is not immoral, for touching your mother's big toe with your little finger is not immoral, and all the rest differs only by degree. But predicates in natural language are almost all vague, and there is no problem in their use; only in formulating the logic that governs them. A vague predicate is usable provided it has clear cases and clear counter-cases.” (van Fraassen, 1980, p. 16)

It is, in other words, not an effective criticism of *any* philosophical position simply to note that it appeals to some vague terms.

The issue, however, is not simply that which theoretical elements count as DIPs is *vague*, but that this is radically *indeterminate*. As Vickers himself (2011) has argued, it is seldom helpful to ask what theoretical posits comprise “the” theory. He makes a similar point when he remarks that:

> “[S]ometimes it is entirely unclear—given the complexities of science and its history—how to identify *the* derivation of some predictive success, or even whether there is such a thing. Posits external to one reconstruction of a derivation might be internal to another reconstruction, such that there is no simple matter of distinguishing DIPs and DEPs.” (Vickers, forthcoming), emphasis in original)

There is, of course, a core set of propositions that will be included in any formulation of “classical electrodynamics”, for instance. But, as stated by the Duhem-Quine thesis (Duhem, 1906/1954; Quine, 1961), it is very rare that *any* empirical results follow deductively from just these core propositions. Derivations of empirical results generally require the postulation of additional assumptions, boundary conditions, etc. By the same logic, even the derivation of *lower-level theoretical posits* from more abstract posits also generally requires additional assumptions. So whether a given posit is related by deductive entailment to any empirical results depends on what other assumptions are included.

Since a DIP is *defined* as a posit that is connected to an important empirical result by truth-preserving (i.e. deductive) inference, a given posit only achieves the status of a DIP because of the role it plays in a larger collection of propositions. DEPs, moreover, may only *fail* to deductively entail the relevant empirical results because the requisite assumptions are not explicitly stated. Indeed, it is a basic theorem of propositional logic that a deductively valid argument can be constructed from a given premise in support of *any* conclusion, provided one is free to select additional further premises at will. So there is always *some* logically possible formulation of a theory under which any given posit counts as a DIP. Any definition of DIPs which appeals to ‘pure’ logical relationships is therefore decidedly unhelpful.

This discussion has, however, drifted into a somewhat uncharitable reading of Vickers. Although his formal definition of a DIP makes relies heavily on the notion of deductive entailment, he also makes it clear that his *intent* is to pick out those "posits which were actually used to derive the impressive result" (Vickers, forthcoming). The partial realist is attempting to explain how particular empirically significant results were actually derived. So she should not be concerned about assumptions that could *conceivably* be considered part of a given derivation, but those that actually historically were part of it. Vickers' criteria for picking out the ECPs of the DIPs only comes into effect once this relatively small set of theoretical assumptions is already specified.

One obvious difficulty with this proposal is that, as argued above, that there will generally be several plausible reconstructions of a given historical derivation. This, however, can be resolved relatively straightforwardly by stipulating that the reconstruction which appeals to the smallest set of assumptions should be regarded as the actual derivation. This is in line with Vickers' approach of attempting to define a minimal core realist commitment; to "at least *some* posits in at least *some* cases". Thus refined, Vickers' position states that the essential elements of an empirically successful theory are the ECPs of the DIPs of the most minimal derivation which was actually used to obtain the empirical result in question.

Vickers' account is similar to interpretation (v) given in the previous section, in that it partly relies on a description of what posits were *actually* appealed to in a successful derivation. Moreover, given the contingency of scientific history, it can always be argued that Vickers' account *might have* categorised as essential more posits than those it does actually so categorise. Unlike interpretation (v), however, his account is not undermined by this contingency, precisely because it adopts such a minimal conception of what elements are essential. Vickers' strategy is simply concede that he has not given the means for identifying *all* the essential elements of a theory, while maintaining that adopting a realist attitude towards those he can identify is still a substantial commitment.

Vickers seems to be arguing that it is better to pick out theoretical elements that one can be relatively sure are covered by the NMA than to provide a more ambitious account which might fall short. Perhaps he is (quite understandably) worried that our reach will exceed our grasp. The problem is that, while the truth of the minimal set of posits picked out by Vickers' criterion does explain the empirical success of the theory in which they are found, it will frequently not provide a *full* explanation. That it counts as an explanation at all is simply because an argument with true premises that has the explanandum as its conclusion is, in general, a kind of explanation. But, by the same principle, the existence of a single black raven is explained by the proposition that all ravens are black. Explanations of this sort are unsatisfactory precisely because the explanatory principles in operation can legitimately be applied with wider scope. Analogously, by failing to apply the NMA more broadly, Vickers' approach, as it were, leaves money on the table. A better, in the sense of more complete, explanation of the empirical success of a theory will (where appropriate) appeal to the truth of theoretical posits which guided scientists in the right direction, albeit not with the force of a deductive argument.

The strongest objection to Vickers' interpretation follows on from the previous objection by pointing the way towards a more complete pattern of explanation. This objection is a variant of that raised against interpretation (iii) (the "convergence point" proposal) above. Suppose that a logical consequence of a posit is an ECP of that posit in respect of some particular derivation. But now suppose, as happens frequently in science, that this posit is involved in several derivations. Which of its consequences we ought to regard as "essentially contributing" may well differ from derivation to derivation. In this case, what is common to all the derivations will be broader than just the part that contributed essentially to the first derivation. Indeed, the common part may be nothing other than the entire posit. Another way of putting this is to recall that the propositions of a theory do not only converge, like the tributaries a river, but may also *diverge* to give rise to various empirical consequences (some of which will turn out to be correct). This information is readily available from the historical record and, more importantly, is explanatorily relevant. So Vickers' account fails to provide complete explanations of the empirical success of scientific theories at least in part insofar as it fails to use this information. In fact, as will be argued in depth in the following section, the most satisfactory account of essentialness depends crucial on identifying these logical ‘divergence points’.

## **10. The empirically successful sub-theory approach**

To introduce this section, let us summarise where things stand. Suppose that some theory of interest produces a statement that a phenomenon will occur. This phenomenological statement is entailed by some theoretical posit (together with boundary conditions, auxiliary assumptions, etc), which is in turn entailed by a ‘higher-level’ posit, and so on. An empirical success occurs when the ‘low-level’ phenomenological claim matches up with what is actually observed. The challenge that faces the partial realist is how ‘far up’ the hierarchy a realist attitude is warranted in light of the successful prediction. The naive realist thinks the whole theory is confirmed. The phenomenological realist thinks that only the low-level phenomenological models are confirmed. Both the structural realist and the proponent of the working posits idea attempt to stake out a middle ground. For the reasons outlined thus far, none of these approaches is without serious defects. The positive account of essentialness offered in this section is an attempt to match the intuitive appeal of some of these other approaches, while staying clear of the obstacles that they have encountered.

When in doubt, it is often helpful to start from first principles. The aim of this chapter is to determine which parts of theories should be counted as essential to their empirical success. This, surely, must depend at least partly on what is counted as empirically successful. As argued in Chapter 2, the most satisfactory account of empirical success is the “unification view” (UV). This view is summarised by the slogan that an empirically successful theory is one which gives rise to more verified empirical content than that required for its construction, and does not give rise to excessive falsified content. The intuitive appeal of this account stems from its compatibility with the NMA. We *expect* that theories which are constructed to fit particular data should account for those data. In contrast, when a theory generates many more empirical truths than are required to construct it, some special explanation seems called for. And, under the NMA, an important part of that explanation is the (partial and/or approximate) truth of the theory.

The positive account of essentialness defended here is called the empirically successful sub-theory account (ESSA). The crucial idea of the ESSA is that the UV of empirical success does not confirm entire theories, but rather provides a means for picking out the confirmed elements *within* theories. To see how this is supposed to work, assume that a scientific theory consists of a set of propositions and their deductive closure. Given this picture, it is always possible to designate a subset of the theory (or ‘sub-theory’). To pick out the essential elements of the theory, start with a sub-theory consisting of statements of its most basic confirmed empirical consequences. These, after all, are the parts of a theory that all parties (even constructive empiricists) agree we should be realists about. Further propositions are added to this sub-theory by a recursive procedure. Consider any theoretical posit not in the sub-theory. If it entails more propositions in the sub-theory than are required to construct it (and does not entail too many negations of these propositions), tag it as confirmed under the UV, and so add it to the sub-theory. Otherwise leave it out. When there are no more theoretical posits to consider in this way, the sub-theory contains the essential elements of the original theory.

This procedure can be illustrated by Fresnel’s theory, a simplified version of which is depicted graphically in Figure 3. The initial sub-theory simply states all the particular observational consequences of the theory that happen to be confirmed. This corresponds to all the boxes on the lowest level of the diagram, minus predictions of the speed of light varying depending on the orientation of the measuring device (labelled “Speed of light”). Now consider the propositions, on the next level up, which express phenomenological laws or models of the type discussed by Cartwright *et al.*. Since “Polarisation” and the “Fresnel equations” both entail more confirmed empirical content than that required to construct them, these propositions are added to the sub-theory. In the next round, the proposition that light consists of “Transverse waves” is added, since this successfully unifies lower-level posits. As expected, however, the posit of a luminiferous ether is not added. The only reason for introducing this posit is to account for the fact that light obeys wave equations. And this posit does not entail any additional verified content from the sub-theory.

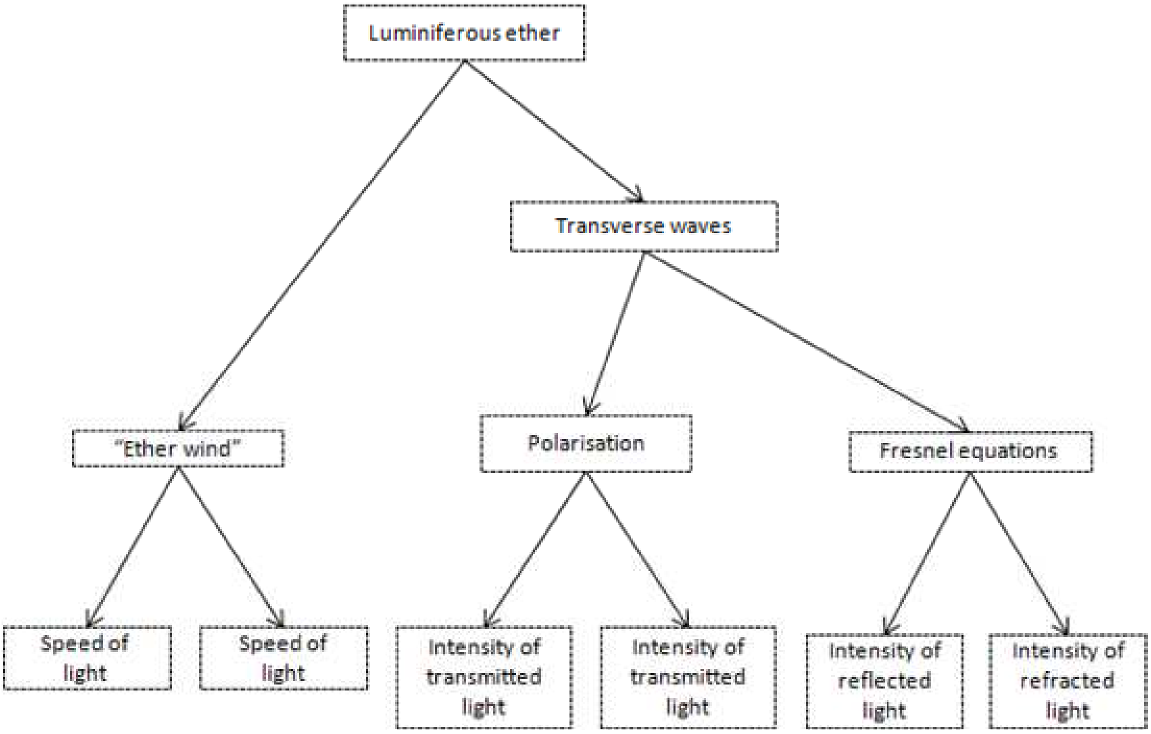
Luminiferous ether
Transverse waves
"Ether wind"
Polarisation
Fresnel equations
Speed of light
Speed of light
Intensity of transmitted light
Intensity of transmitted light
Intensity of reflected light
Intensity of refracted light

**Figure 3.** A diagram showing some elements of Fresnel's theory of light. Posits are represented by boxes, and deductive relations by arrows between them.

How well the ESSA measures up against some of the desiderata for an account of essentialness articulated above, as well as its additional consequences and potential difficulties, is best explored by contrasting it with the alternative views discussed in this chapter. One major difference between the ESSA (and the working posits idea, incidentally), on the one hand, and entity/phenomenological realism and more restrictive versions of structural realism, on the other, is that the former does not make any *general* claims about which 'level' of a theory contains the essential elements. In some specific cases, the algorithm sketched out above will terminate with statements of phenomenological laws. The London model of superconductivity may well be just such a case. In other cases, such as Fresnel’s theory, there is a stronger case for adopting a ‘realist attitude’ about ‘mid-level’ theoretical generalisations. And, in yet other cases, highly abstract ‘metaphysical’ claims may turn out to be essential under the ESSA. For instance, if, counterfactually, the ether wind had turned out to be a real phenomenon the ESSA would *demand* a realist attitude towards the luminiferous ether.

There are also, however, crucial differences between the ESSA and the notion of working posits. Whereas neither attempts to make an *a priori* distinction between posits at different ‘levels’, the latter (especially in Vickers’ interpretation, but also in Psillos’) does distinguish between posits that are “internal” versus “external” to a derivation. Notice, moreover, that Vickers’ distinction between DIPs and DEPs mirrors Cartwright’s distinction between the role played by phenomenological laws and fundamental laws. Recall that, in Cartwright’s view, particular empirical models are not deductively derivable from fundamental laws, although their construction is “guided” by such laws. This is sufficiently similar to Vickers’ description of DEPs as posits that are not connected to empirical results by a truth-preserving derivation, but serve to “inspire” such a derivation, that the two arguments may be addressed together.

Several arguments have already been offered to the effect that the distinction suggested by Cartwright and Vickers does not justify a differential ‘realist attitude’ towards the posits concerned. The first set of arguments was provided in section 9 of this chapter, and aimed to demonstrate that the proposed distinction cannot bear the weight placed upon it. However, although no firm distinction between ‘internal’ and ‘external’ posits is likely to be tenable, there is certainly, as argued in section 4 of this chapter, some “*spectrum*” of cases. At one end of this spectrum lie cases where a theoretical posit, together with a few ‘natural’ assumptions and boundary conditions, entails the required empirical model. At the other end lie cases where the requisite auxiliary assumptions are far less ‘natural’ and, indeed, are likely only to be invoked so that high-level theory can be offered as a *post hoc* explanation of some successful empirical model. Cartwright, by focussing on cases where the deductive link is more tenuous, gives the misleading impression that there are no cases where it is more straightforward. Presumably, Vickers would assent to this criticism.

Nevertheless, it seems that Vickers would agree with Cartwright regarding cases in the ‘middle’ of the spectrum, where high-level theoretical posits can be said merely to “guide” the formation of a model. Cartwright is willing to infer that a particular entity exists if this is the best *causal* explanation of a given empirical regularity. As argued in section 4, however, this type of reasoning should apply equally well to *types* of causal interaction, and thus to posits of high-level theory which describe such interactions. Neither Vickers nor Cartwright adopt this position because both focus on *individual* derivations of empirical results or models. From such a perspective, the postulation of a causal interaction which is logically far-removed from the particular empirical result seems quite superfluous. However, when the focus turns to a *collection* of derivations, and it is apparent that the same abstract posit serves a common inferential role in all of them, the criterion of unifying power may quite justifiably lead one to regard this posit with a realist attitude.

This is not, of course, to claim that *any* “unifying” hypothesis ought to be regarded as probably true by the NMA – the connection between hypothesis and the various empirical results may nevertheless be so tenuous and *post hoc* that the inference does not follow. As emphasised in Chapter 2.10, there are several factors which together comprise the unifying power of a particular theoretical posit. At stake in the present discussion are at least two of these factors, namely the ‘breadth’ of unity and the ‘strength’ of the logical connection between hypothesis and the empirical results.

Since both of these factors vary on a gradient, there will always be cases where it is unclear to what degree a realist attitude is warranted. Nevertheless, contra Cartwright and Vickers, there will also be cases where a realist attitude is warranted towards a hypothesis that merely “inspires” empirical results simply because it has this relationship to many and varied such results.

This discussion of unity raises another advantage that the ESSA has over the working posits idea. As was made apparent in the previous two sections, one of the major difficulties in interpreting the working posits idea is the tension between causal/historical and logical interpretations. Either demand, when expressed in its purest form, leads to disaster. Vickers has made an extremely impressive attempt at developing a stable compromise in response to these competing demands but, as argued in the previous section, this ultimately fails. The ESSA, unlike the working posits account, is able to offer a purely logical criterion of essentialness without catastrophe. This is because it focuses on a different logical relationship, namely unification as opposed to derivability. Or, applying the metaphor given in the previous section, the ESSA can be said to focus on *divergence* rather than *convergence* points in the 'stream' of logical derivation.

I turn now to the question of whether the ESSA provides the means to *explain* why a given theory produces accurate empirical claims. In answering this question, it is worth being clear on exactly what is meant by "empirical success". Considering a *particular* empirical success, i.e. one instance of an empirical prediction of the theory matching up with observation, it is not obvious that the ESSA is able to furnish a satisfactory explanation. The ESSA claims as essential those propositions of a theory which, by deductive or less strict forms of inference, unify verified lower-level propositions of the theory. But it is not at all clear that the occurrence of a particular empirical result is explained by the truth of proposition which unifies it with other empirical results. Indeed, since the proposition may count as unifying under the UV even if the relationship between it and all the empirical results it unifies falls well short of deductive entailment, the truth of this proposition does not explain this empirical success even by the logical derivability standard dismissed as inadequate above. The truth of a unifying hypothesis does not, in other words, explain each instance of successful empirical prediction individually. But it can explain the *overall* empirical success of a theory, i.e. its ability to correctly state the results of many and varied empirical observations. The counterfactual formulation makes the intuitive appeal of this statement more obvious: If the unifying hypothesis were not true, then we would not expect to see such broad empirical success. So the truth of a unifying hypothesis does explain the broad empirical success of a theory in which it is found.

This brings us to a final point. The ESSA is motivated by the unification view of empirical success and, as argued in Chapter 2.9, the extent to which a theory is unifying can only be determined when considered in light of the overall construction of a theory “from nothing”. By the same argument, then, the degree to which a particular *theoretical posit* unifies must also be understood within this larger context. Figure 4 is therefore a more complete (though not exhaustive) picture of Fresnel’s wave theory of light, using the additional details concerning the construction of this theory discussed in Chapter 2.9. Notice that all the information found in Figure 3 can be found embedded in this larger picture, although some details have been omitted for clarity.

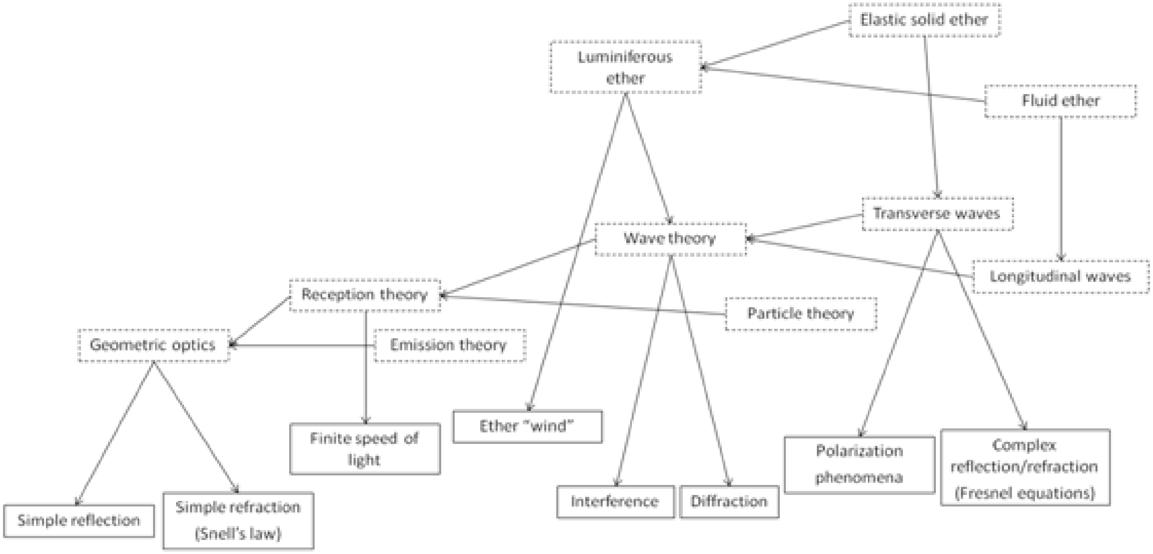

**Figure 4**. Theories of light up to Fresnel. Logical relationships between posits are shown by arrows, with the proposition at the foot of an arrow entailing that at its head. Dashed boxes denote theoretical posits, and solid boxes represent descriptions of empirical phenomena.

Every particular instance of a theory entails the more general version from which it is derived. So both the wave and particle theories of light, for instance, entail the reception theory *and* all its associated consequences. Reading off the figure, it is apparent, for instance, that both wave and particle theories have the consequence that the speed of light is finite. When considering whether a given posit ought to be accepted under the unification criterion, one should therefore tally up *all* empirical consequences it gives rise to, and assess whether these are greater than what would be required to deduce this posit “from nothing”. Equivalently, but more practically, we can determine whether each *step* in the (actual or hypothetical) construction of the theory increases unifying power. For instance, the non-specific wave theory of light correctly predicts interference and diffraction phenomena, which is more additional confirmed empirical content than that required to derive it from the more general reception theory.

There remain legitimate grounds for criticism of the ESSA, as it has been presented above. The most important criticisms emerge from the simplifications that have been introduced in order to give the neat ‘tree’ picture of scientific theories exemplified by Figures 3 and 4. The first of these simplifications is that, in characterising a theory as a set of propositions and entailment relations, a “syntactic” view of theories has been presupposed. This is opposed to a “semantic” view, which understands theories to be characterised by families of models. Although the semantic view is currently the prevailing fashion, I tend towards the view (following Worrall, 1984; and French, 2009) that these two views simply express different sides of the same coin. In brief, the argument for this view is that a given class of model can be characterised by the propositions that they satisfy, and vice versa. So, although the syntactic view is used here for clarity of exposition, it is plausible that the same basic arguments can be made, mutatis mutandis, in the semantic framework.

The second major simplification is that the ‘tree’ picture also seems to imply that all the important intratheoretical relations are straightforwardly deductive. In light of the arguments provided in Chapter 2.10 and in this chapter against the view that deductive relations have a qualitatively different epistemic status, this is obviously not the case. The Fresnel theory, as an example, conforms to the deductive model relatively well. But, in general, the simple tree picture must be made be complicated. Posits which merely “influence” or “guide” the proposal of empirical statements or lower-level theoretical posits should be included, with dotted arrows showing these ‘weaker’ logical relations. The essentialness of a hypothesis is thus understood as the product of at least two factors, namely the ‘breadth’ of the empirical results that follow from the hypothesis and the ‘strength’ of the logical connection between the hypothesis and these results.

A picture of scientific theories that includes non-deductive relations immediately poses two problems, however. Firstly, since the unifying power of a posit is held to depend partly on the ‘strength’ of its logical connection to other posits, there must be some account of how this strength is to be determined. Secondly, because it is not at all clear how to decide whether a particular model conforms to a set of propositions that includes non-deductive relations, this picture seems incompatible with the semantic view of theories after all.

The most promising solution to both problems is simply to ‘fill in’ the assumptions required to turn these weaker relations into deductive relations. This immediately resolves the issue of compatibility with the semantic view. And the strength of a given non-deductive relation can quite straightforwardly be determined on the assumption that it is inversely dependent both on the number of additional assumptions that must be invoked for it to be rendered deductively valid and how ‘natural’ these are. Of course, very significant questions remain over how we are to categorise assumptions as more or less “natural”. These questions are, of course, an endless source of problems to philosophers of science, and they will not be tackled here. The intent of this admittedly rough discussion has been simply to give a sense of how the ESSA could be applied to the details of actual cases. A more concrete demonstration will come *via* the consideration of actual case studies in the following chapters.

## **11. Chapter summary**

The goal for this chapter has been to give a satisfactory account of which elements of empirically successful theories we ought to be realists about, assuming the validity of the no-miracles argument. We have sought, in other words, to give a general story about which elements of theories we ought to regard as ‘essential’ to their success. To this end, I have examined five broad approaches to this question which are already present in the literature, namely: referential continuity approaches; entity or phenomenological realism; structural realism; semi-realism; and the working posits idea.

Although several of the other accounts of essentialness draw upon the notion of referential continuity, in section 2 I considered a relatively ‘pure’ form of the notion, which regards any two theoretical terms as co-referring just in case they are both used when causally interacting with the same actual entity. However, since any two scientists who attempt to describe a given empirical phenomenon will undergo the same causal interactions, the demand for continuity between theories will be far too ‘easily’ satisfied. This is not to mention the fact that an account of essentialness constructed along these lines will be essentially backwards-looking and so relatively uninteresting. To give a non-trivial account of essentialness, therefore, requires that we designate some parts of the theory which attribute *properties* (including relational properties) to entities as being essential to the theory’s success. Once we have such an account in hand, we can make the *post hoc* judgement that the parts of the theory so designated *refer*. But this is a sort of metaphysical ‘optional extra’.

In section 4, I discussed somewhat different forms of entity realism due to Hacking and Cartwright, as well as a development of Cartwright’s views which has been termed “phenomenological realism”. What all these views have in common is that they regard as essential just the most basic empirical laws or models of a theory. Equivalently, we can say that these accounts regard as essential the most basic properties by which we identify particular entities (the “detection properties” of these entities, to use Chakravartty’s term). My objections to this type of view were twofold. Firstly, I argued that there are many cases where the statement of low-level empirical laws or basic properties follows quite directly from higher-level theoretical posits. The phenomenological realist's arguments for regarding these theoretical posits as inessential therefore fail in these cases. Secondly, even in cases where the connection between abstract posits and empirical laws is more tenuous, the truth of the abstract posit might nevertheless feature in the best explanation of the success of the empirical laws and so be worthy of a realist attitude. As I argued in section 8, this is especially the case where a high-level posit is involved in deriving *several* accurate empirical laws.

In section 5, I considered structural realism, which regards as essential only the *structure* of a theory, i.e. the overall set of relations posited to exist between entities. The decisive argument against structural realism has been articulated by Chakravartty and Stanford. It notes simply that the notion of 'structure', defined as above, is overly inclusive and thus fails to exclude paradigm cases of 'inessential' theoretical posits such as the luminiferous ether. A structural description of a theory, although it may help to clarify the logical relationships between elements, therefore must be supplemented by some *additional* criterion of essentialness. And so it is clearly not an adequate account of essentialness in its own right.

I have devoted a great deal of space – sections 7-9 – to a discussion of the working posits idea, since it is currently so influential in the literature. I have pointed out, however, that this basic idea is in fact open to several different interpretations. The primary tension animating the different interpretations is that between causal and logical accounts. A 'pure' causal account is narrowly focused on which theoretical beliefs were *actually involved* in bringing scientists to the explanatory or predictive feats that are our paradigms of scientific achievement. A 'pure' logical account, in contrast, is concerned with which theoretical propositions are *logically required* to achieve these empirical successes. Either type of pure account results in unacceptable consequences. Vickers attempts to achieve a stable compromise between these two by arguing that the essential elements are the "essentially contributing parts" of "derivation-internal posits". I have argued, however, that has not succeeded resolving the underlying tension between causal and logical considerations.

Finally, I offered my own the empirically successful sub-theory account. This suggests that the essential elements of a theory are just those which successfully account for more empirical phenomena (or lower-level theoretical posits) than that required to construct them. This has major advantages over all the other accounts outlined above. Both referential approaches and an inclusive version of structural realism are insufficiently selective, in their own ways, and thus trivial. The ESSA, in contrast, allows us to say prospectively which elements of a theory are or are not essential. Note, however, that the ESSA can very naturally be applied to a *structural interpretation* of a theory, thus giving an ‘essential structure’ account. The theoretical terms picked out by the ESSA, moreover, can be interpreted as ‘referring’ or ‘representing reality’, although I do not wish here to be drawn into metaphysical debates regarding whether these interpretations are well-founded. Unlike phenomenological realism and certain interpretations of the working posits idea, the ESSA does not attempt to argue that only theoretical posits which deductively entail empirical results are ‘essentially involved’ in the derivation of these results. The most important reason for this is that the truth of a hypothesis which has some inferential, but not necessarily truth-preserving, link to a set of empirical claims can be a part of the best explanation of why these claims successfully describe the phenomena. So, because the ESSA connects neatly to the explanatory considerations which motivate scientific realism, and because its relevant competitors suffer from major weaknesses, the ESSA is the best account of which parts of scientific theories ought to be regarded as essential to their empirical successes.

# **Chapter 4. Rationality and theory choice**

## **1. Chapter overview**

The major aim of this chapter is to address the long-standing question concerning rationality in theory choice in light of the accounts of scientific realism developed above. The stage is set by Kuhn’s claims that any assessment of a “paradigm” is dependent on epistemic values that are themselves paradigm-dependent. This suggests that there are not any objective, rational standards for theory choice.

Section 2 surveys four broad views on the role of objective standards in theory choice. The criterion of rationality under “Whig history” is acceptance of whichever theory is most similar to the *current* best theory. But this dependence on hindsight means that Whig history fails to offer normative criteria of judgement that might have been applied by actual historical scientists, who of course do not know what theories future scientists will accept. The “Strong Programme”, in contrast, states that theory choice is largely driven by non-rational factors, such as political sentiment. It is implausible, however, that rational factors (even paradigm-dependent ones) play *no* role in theory choice. The two views that are considered more seriously in this chapter are “Kuhnian history” and “rationalism”. Kuhnian history states that there are objective criteria for theory choice, but that these do not fully determine what is rational in most cases. Non-rational factors therefore play some role. Rationalism, in contrast, states that it is possible in many or most cases to determine clearly which of the theories on offer it is rational to accept.

A particular version of rationalism, named “partial rationalism”, is described in section 3. It states that, during episodes of theory change, scientists are rational to accept those *parts* of theories responsible for empirical successes (and have largely done so). This position follows relatively directly from partial realism, since to state that certain parts of a theory accurately represent the world is *a fortiori* to regard to them as rationally acceptable. It also follows from deflationary realism, since a belief in the general continuity of science generates a practical commitment to preserving the essential elements of current theories. An interesting consequence of partial rationalism is that, in some cases, it demands (and predicts) the creation of a *synthesis* containing the essential parts of competing theories, rather than a choice between them. This is in direct opposition to the Kuhnian approach to theory choice.

Sections 4-6 provide detailed case studies in support of the partial rationalist position. Section 4 focuses on the Copernican revolution. Following Worrall and contra Kuhn, it is argued that the Copernican model of the universe was indeed rationally superior to the competing Ptolemaic system, and this rational superiority explains its eventual dominance in the community. Sections 5 and 6 focus on cases of theoretical synthesis. Section 5 examines the “neo-Darwinian revolution”, in which the concept of natural selection from Darwinian theory was combined with Mendelian particulate genetics to yield a unified account of genetic change in biological populations. Section 6 examines the “prion revolution” in molecular biology. Prior to this, it had been argued that biological replication always involved the transfer of sequence information in the form of nucleic acids. The prion theory postulated another form of biological information, namely protein conformational information, which could also be replicated.

## 2. Rationality and theory choice

A useful starting point for any discussion of rationality in science is Kuhn’s (1962/1996) *Structure of Scientific Revolutions*. The central thesis of this book is that the scientific research taking place in any given discipline at any given time is framed and guided by a particular “constellation of group commitments”, which Kuhn refers to as a “paradigm”. There are several elements to a paradigm, including a set of shared examples of successful scientific practice upon which further scientific research is to be based. Crucially for the present discussion, a paradigm also includes shared “values” which govern what count as pressing scientific problems, and what count as solutions to these problems. Because different paradigms involve commitments to different epistemic values, disputes *between* competing paradigms tend to involve circular arguments. The practice of a particular group of scientists cannot be assessed by reference to some external set of epistemic standards, since each group judges the value of scientific achievements precisely according to the values of the paradigm it accepts.

A seemingly inevitable consequence of this picture is that episodes of theory-change in the history of science cannot be considered "rational" by any objective standard. Kuhn makes this point quite clearly by arguing that these episodes are analogous to political revolutions. It is worth quoting the relevant passage at length:

> "At that point [in a political revolution] the society is divided into competing camps or parties, one seeking to defend the old institutional constellation, the others seeking to institute some new one. And, once that polarization has occurred, political recourse fails. Because they differ about the institutional matrix within which political change is to be achieved and evaluated, because they acknowledge no supra-institutional framework for the adjudication of revolutionary difference, the parties to a revolutionary conflict must finally resort to the techniques of mass persuasion, often including force... The remainder of this essay aims to demonstrate that the historical study of paradigm change reveals very similar characteristics in the evolution of the sciences. Like the choice between competing political institutions, that between competing paradigms proves to be a choice between incompatible modes of community life. Because it has that character, the choice is not and cannot be determined merely by the evaluative procedures characteristic of normal science, for these depend in part upon a particular paradigm, and that paradigm is at issue." (*ibid.*, pp. 93-94)

Since all rational standards for determining which of two scientific paradigms is superior are themselves paradigm-dependent, the choice between the two will inevitably depend partly on non-rational factors (though in the sciences, thankfully, the resort to force is less common than in the political realm).

These remarks are highly suggestive, but do not give a completely clear picture of whether, and to what extent, objective rational factors play a role in scientific reasoning. The remainder of this section is therefore a survey of four broad accounts of scientific rationality in the context of theory choice. These are: (i) Whig history; (ii) the Strong Programme; (iii) the (mature) Kuhnian view; and (iv) rationalism. I will argue that the first two approaches are untenable, leaving only the latter two as viable options. Following on from this, in section 3, one particular version of rationalism, namely partial rationalism, will be articulated. Partial rationalism and Kuhnian history will then be compared and assessed by reference to particular historical case studies.

Butterfield's (1931/1965)) aim in coining the term "Whig history", in the context of political history, was to diagnose and criticise the pernicious tendency (as opposed to a consciously-endorsed position) among historians to project their own concerns and beliefs anachronistically onto actors in the past. The basic idea of Butterfield's argument is simply that historical actors behaved as they did for reasons that may have differed markedly from our own. Their actions may, in the long run, have had the effect of bringing about consequences which we ourselves value. It is, however, very frequently just historically inaccurate to think of them as acting with the intention of producing these consequences or, if they did aim at these consequences, to think that they did so with the same motivations that we have in achieving them. What goes for descriptive accuracy also goes for normative evaluations – it is frequently a mistake to regard some actor as a hero (villain) because she acted in such a way as to bring about a goal we ourselves value (abhor).

In the history of science specifically (Jardine, 2003), "Whiggishness" results in the mistaken ascription of epistemic virtue to historical scientists who supported theories that are approximately correct from the perspective of our current theories. So, in instances of theory change, whichever scientists supported the theory closest to our own is automatically described as rational. And, since newly-proposed theories have typically been closer to ours than their pre-existing rivals, this usually amounts to a judgement in favour of the proponents of novel theories. Their opponents, in contrast, are usually viewed either as irrational – mistaken about some factual matter, or in the grip of some cognitive delusion – or acting from motives that are not (scientifically) rational, perhaps to preserve their positions of power and social influence.

This asymmetric treatment of proponents versus opponents of novel scientific theories is extremely difficult to justify as an *a priori* attitude. While we are certainly entitled to judge that past actors were *incorrect* in their scientific judgements, it seems entirely unwarranted to judge that they were *irrational* simply on that basis. The task of any criterion of rationality is to warrant certain inferences against a background of beliefs and observations already accepted by the actor concerned. A Whiggish criterion of rationality fails in this task, as it takes as its starting point beliefs and observations available to the later historian, but not to the historical actor being judged. We may, of course, legitimately conclude that one or other party is less rational than another, but this can only be done *after* examination of the information available to each of them.

Notice, incidentally, that this criticism of Whig history mirrors very closely the criticism found in Chapter 3.3 of retrospective accounts of essentialness. This similarity emerges because assessing whether some (part of a) theory is responsible for empirical success is logically related to assessing whether it is rational to accept this theory. The parallels between delimiting realist commitment and rational criteria for theory choice are drawn out in more detail in the following section of this chapter.

The “asymmetry” of Whig history is tackled directly by the second approach to the history of science, namely the “Strong Programme”, most famously associated with Barnes and Bloor (1982; 1996; Bloor, 1976). The primary doctrine of the Strong Programme is the so-called “equivalence postulate” or “symmetry thesis” which states that:

> "[A]ll beliefs are on par with another with respect to the causes of their credibility… This means that, regardless of whether the sociologist evaluates a belief as true or rational, or as false and irrational, he must search for the causes of its credibility." (Barnes & Bloor, 1982, p. 23)

While strongly worded, it seems that few could object to this statement, in light of the obvious flaws of Whig history.

In fact, however, the Strong Programme is controversial because of the emphasis that its adherents have placed on non-rational factors in influencing the course of scientific history. For example, in one highly-regarded product of the movement, Shapin and Schaffer's (1985) *Leviathan and the Air Pump*, the authors examine the struggle between the competing scientific approaches of Boyle and Hobbes in Restoration-era England. They argue that it was ultimately because the complex of ideas associated with Boyle was *politically* successful that his scientific ideas triumphed. This conclusion is stated explicitly in the closing chapter of the book:

> "We have attempted to show … that the contest among alternative forms of life and their characteristic forms of intellectual product depends upon the political success of the various candidates in insinuating themselves into the activities of other institutions and other interest groups. He who has the most, and the most powerful, allies wins." (*ibid*, p. 342)

So, whereas Whig history casts one side of a scientific dispute as rational, and its opponents as irrational or at least non-rational, the Strong Programme refuses to attribute rationality to *either* side. Political ideologies, personal interests and other 'sociological' factors essentially decide the outcome of these disputes. Another way of putting this is that proponents of the Strong Programme interpret the analogy offered by Kuhn in the previous section very literally – there is no substantive difference in the type of discourse that occurs in political versus scientific revolutions.

The Strong Programme's initial diagnosis of the error in Whig history is entirely correct. The truth of some scientific theory (from the perspective of a later theory) is, indeed, not by itself a sound explanation for the success of that theory at an earlier point in history[17]. But this only excludes one possible account of rationality. By weighing in against rationalist accounts of science in general, the proponent of the Strong Programme is therefore tacitly endorsing the same mistaken account of rationality as the Whig! In general, an explanation by reference to truth is not the same as an explanation by reference to rationality. To say that the proponents of some theory were "rational" is simply to say that they had good arguments in their favour. And a theory having good arguments in its favour is a perfectly acceptable explanation for its success. It is surely uncontroversial to state that human beings do, at least some of the time, behave in certain ways because they were convinced by argument to behave in those ways. Moreover, given the large amount of time that scientists spend presenting arguments for and against theories, it is implausible to claim that all this activity is merely a sideshow to the political manoeuvres which actually decide the issues. So considerations of rationality must play *some* role in episodes of theory change. The two remaining views considered in this section differ over exactly what this role is.

The third approach to the history of science emerges from the later work of Thomas Kuhn (1962/1996, see especially the 1969 postscript; 1977). Here, Kuhn appears to retreat somewhat from his apparent wholesale dismissal of objective values in the first edition of *Structure.* Instead, he articulates a view about the rationality of scientific revolutions that is a hybrid of rationalist and anti-rationalist accounts. He argues that scientists apply "values" or "objective factors" in choosing between scientific theories. These include the *accuracy* with which the theory matches phenomena; the theory's *consistency*, both internal and with reference to other theories; its *scope*; its *simplicity*; and how *fruitful* it is in generating further research.

[17] This, of course, is virtually the same argument made by Stanford *et al.* against retrospective versions of partial realism (see Chapter 3.3).

These criteria, however, may conflict in particular cases, and even the application of a single criterion can lead to conflicting judgements. One theory may, for instance, be highly accurate in accounting for a given class of phenomena, but less accurate than its rival in another domain. Providing a fully objective and deterministic “algorithm” for theory choice would require some additional standards for applying these virtues, and for weighing the relative importance of virtues when they conflict. But, argues Kuhn, there are no such objective higher-order standards. The demands of rationality are therefore indeterminate. Kuhn believes that it is non-rational factors which ultimately ‘close the gap’ between rationality and theory choice. Scientists are to some irreducible degree guided by aesthetic or otherwise idiosyncratic personal preferences, social and political pressures, etc. when deciding the extent to which a given theory satisfies the theoretical virtues. And all of this counts as rational for Kuhn, so long as these other factors do no more than narrow down the range of possible answers that are compatible with the basic values of rationality.

The fourth and final approach to the history of science is what might be called “rationalist”. This view is somewhat aligned with Kuhn’s account, in that it suggests that there are scientific values or “virtues” which theories might fulfil and that fulfilment of these virtues makes it rational to choose them over their competitors. However, where Kuhn argues forcefully against the notion of an “algorithm” that will uniquely choose the best theory amongst several contenders, rationalism claims that there is a hierarchy among the virtues, or some other general principle that will at least sometimes render accepting one theory, on balance, *more rational* than accepting its competitor.

It should be emphasised that rationalism involves two distinct commitments, one normative and one descriptive. Normatively, it claims that there is some objective rational standard or algorithm for deciding between scientific theories. The descriptive claim is that, in most or all actual instances of theory change in the history of science, the theory that eventually became dominant did so because it was superior by these standards. There are different forms of rationalism corresponding to different views on which set of epistemic standards is to be normatively endorsed. In the following section, one particularly promising formulation will be articulated and defended.

## **3. Partial rationalism**

Partial rationalism ultimately originates from a criticism of Kuhn's view given by Worrall:

> "[Kuhn's] argument against the objectivist nonetheless goes through *only* if we accept the initial assumption that the objectivist can do no better than supply a "laundry list" of objective factors, and is therefore left entirely without recourse when two factors from the list pull in opposite directions. But I know of no objectivist [i.e. rationalist] who would accept Kuhn's list as it stands and none who would be happy to leave *any* such list unstructured. For example, for Duhem, Poincaré, Lakatos, and many others, there is a *basic* criterion: that of predictive empirical success. When this criterion is properly understood, it informs most of those on Kuhn's list." (Worrall, 1990, pp. 333-4)

In the remainder of this section, this quotation will be interpreted in a way which may well, but does not necessarily, accurately represent what Worrall himself was attempting to convey. Firstly, "predictive empirical success", instead of being understood by reference to the UN account, will be read as "success under the UV" (as discussed in Chapter 2). Secondly, the notion of "theory confirmation" in the quotation above will be interpreted in the light of scientific realism. That is, it will be assumed that theories which enjoy predictive empirical success should be understood as *true* (albeit only approximately, at the structural level, etc.). Given this commitment, it is not at all surprising that empirical success should be the basic theoretical virtue in cases of theory choice. Empirical success (of the appropriate sort) is, by hypothesis, sufficient to underwrite the belief that a theory is (partly and/or approximately) true, and the standard for believing that a theory is *true* is surely stronger than that for merely preferring it over rival theories!

It is worth giving a general idea of how Worrall's claim that the criterion of UN predictive success "informs most of those on Kuhn's list" ought to be interpreted. Recall that the UV states that a theory (or theoretical posit) is confirmed just in case it gives rise to more accurate statements about empirical phenomena than that required to construct it (and does not give rise to too many *inaccurate* empirical statements). From this definition, it is immediately apparent that the virtue of (i) *accuracy* is easily incorporated. Similarly, a unifying posit must have relatively broad (ii) *scope.* This is clearer when stated negatively – a hypothesis which entails only a single set of empirical phenomena by definition fails to unify phenomena. As argued in detail in Chapter 2, the criterion of unifying power is also intimately related to the virtue of (iii) *simplicity*. Briefly, the more simple a theory is, the fewer assumptions or free parameters it contains, and thus the fewer verified empirical results are required to construct it.

The virtue of fruitfulness does not fit quite so neatly, and so warrants somewhat more extensive discussion. Worrall expresses his view on the topic as follows:

> ""Fruitfulness" too is intimately connected to predictive success. A general theoretical approach... shows its fruitfulness by supplying ideas for developing specific theories *independently of empirical results*. Such an approach will be judged barren... only when all these ideas have been tried *without predictive success*; and hence the approach has been reduced to tagging along *behind* the empirical data, always accommodating it *post hoc* rather than predicting it in advance." (Worrall, 1990, pp. 334, emphasis in original)

Notice, firstly, that the conclusion of Worrall's argument, as it stands, states merely that fruitfulness and empirical success are regularly *associated*, not that the former is partly *constitutive* of the latter. This is perhaps not surprising since, from the realist perspective, fruitfulness has at best a very weak claim to being an *intrinsic* theoretical virtue. The hypothetical ‘final theory of everything’ (or even of some particular empirical domain) cannot be improved upon, and so poses no further research questions.

A second thing to notice is that Worrall appeals to a novel prediction account of empirical success, not the UV. Nevertheless, there is at least one argument for thinking that fruitfulness is associated with empirical success under the UV. This argument links fruitfulness strongly to the idea of scope which, as pointed out above, is partly constitutive of unifying power. The argument starts from the (uncontroversial) claim that we are nowhere close to a ‘final theory’ in any given scientific domain (if this is even a possibility). In this context, a fruitful theory is one which poses many interesting research problems, i.e. which makes many specific, but as-yet unverified, empirical claims. But a theory with wide scope, i.e. which gives rise to a wide variety of verified empirical results, is very likely also a theory which gives rise to as-yet *unverified* empirical results. So a unifying theory is therefore likely to be fruitful.

Finally, Worrall claims that one item on Kuhn’s list is not in fact a virtue at all, namely consistency with existing theories:

> “It is surely a *virtue* in a theory, rather than a vice, if it clashes with some well-entrenched claim – *provided* that there is strong evidence for the theory in the form of predictive empirical success... Scientists will, no doubt correctly, downgrade... new theories that clash with well-established ones – but *only* when there is no independent evidence for the new theory.” (Worrall, 1990, p. 335, emphasis in original)

So, just as one might expect from Worrall’s realist motivations, empirical success (understood, on the present reading, under the UV account) is regarded as the *sole* genuine theoretical virtue. Other putative virtues are valuable only insofar as they represent aspects of, or are reducible to, unification.

The advantage, from the rationalist perspective, of having a single basic theoretical virtue is that Kuhn's arguments for indeterminacy are rendered inapplicable. All theories are, as it were, measured on the same scale. Moreover, although some might dispute the particular account of empirical success favoured in this thesis, each of the candidates discussed in Chapter 2 is at least relatively clearly defined. There is thus far less scope for the situation where each of several competing theories is superior on differing interpretations of a single theoretical virtue.

However, while this form of rationalism avoids certain types of indeterminacy, it threatens to create others. Consider, for instance, the scenario where two or more mutually exclusive theories are *both* empirically successful to some degree. To choose one of these theories is to discard the other. But, by hypothesis, the fact that each is empirically successful implies that it is at least partially and/or approximately true. So to discard one of these theories would seem to require rejecting some true statements.

The dilemma created by this situation strongly mirrors that highlighted by the pessimistic meta-induction, as discussed in Chapter 1. And, similarly, this dilemma is resolved by conceding that the no-miracles argument for realism, at best, supports only *partial* realism. So, to the extent that realism entails a commitment to rationalism, the form of rationalism it entails what might be called "partial rationalism". Partial rationalism states that, during episodes of theory change, scientists are rational to accept those *parts* of theories responsible for empirical successes (and have largely done so). And, while theories taken as a whole may be mutually incompatible, it is much less likely that the 'essential' parts of each of the theories concerned will form an inconsistent set.

There are at least two types of case of *prima facie* theoretical incompatibility which the partial rationalist must address. The first is that in which the older of two competing theories, although strictly incompatible with the newer theory, is nevertheless *approximately* true by lights of the latter. For instance, although it has been superseded by relativity, Newton’s theory of motion gives quantitatively accurate predictions in cases where all the objects concerned have low speeds relative to that of light. Although the new theory is more empirically successful than the older, and so is chosen over it, the empirically successful parts of the older theory are nevertheless retained. Therefore no true statements are rejected. One example of this sort of case is illustrated in section 4 of this chapter.

This first type of case is the usual subject of case studies by partial realists. There is also, however, another type of case, where no such relation of approximate preservation exists between the competing theories. Here no choice between theories is possible without outright rejection of some essentially contributing parts of one or other theory. The partial rationalist does have another point to make, however, namely that in these sorts of cases the Kuhnian emphasis on a *choice* between theories is in fact extremely misguided. In these cases, partial rationalism demands instead that we accept *all* the parts of the competing theories which are essential to their respective empirical successes. It demands, in other words, *synthesis* rather than a choice. This somewhat abstract point will be illustrated with examples, in sections 5 and 6 of this chapter.

Before moving on to illustrative case studies, it is worth briefly outlining how the remarks made in this chapter and elsewhere in this thesis amount to *arguments* for partial rationalism. The first argument is *a priori*. It begins with the arguments for partial realism, including the version of the no-miracles argument which regards unifying power as a marker of truth. If it is accepted that partial realism is well-founded, partial rationalism then follows readily as a consequence, as demonstrated above. The second argument is premised instead on the truth of *deflationary* realism. The descriptive claim of partial rationalism follows straightforwardly from DR. Moreover, as argued in Chapter 1.8, a belief in the actual continuity of science generates, by pragmatic reasoning, a normative commitment to preserving the essential elements of currently successful theories in proposed successor theories.

The third argument is suggested by Worrall in the quote above, where he states that: “When this criterion [of predictive empirical success] is properly understood, it informs most of those on Kuhn’s list.” This can be interpreted as an argument for partial rationalism, as follows. Suppose it is accepted that each of Kuhn’s theoretical virtues has at least some intuitive appeal. Now suppose that each of these virtues is quite naturally interpreted as a component or aspect of one basic virtue. This basic virtue is also intuitively appealing in its own right. By parsimony, it is then reasonable to conclude that these other virtues are not normatively desirable in their own right, but that their intuitive appeal can be attributed solely to their relationship to the basic virtue. Thus, if the arguments which purport to demonstrate that Kuhn’s theoretical virtues (with the exception of consistency) are reducible to the criterion of empirical success are well-taken, partial rationalism is supported.

The fourth argument for partial rationalism replicates a pattern of argument found frequently in philosophy of science. This pattern of argument begins from the premise (i) that scientists’ actual reasoning at least closely approximates what is normatively desirable. It then states (ii) that scientists actually do reason as we would expect if they endorsed the methodological norm that is being defended. Finally, it concludes (iii) that application of this norm is, in fact, normatively desirable. Issue may be taken with this type of argument, especially with premise (i). Nevertheless, if it is in general sound, an argument for partial rationalism can be constructed on the basis of historical examples which demonstrate that scientists do at least some of the time reason as if they accepted partial rationalism. The case studies which follow in this chapter can therefore be read as underpinning such an argument.

The fifth argument for partial rationalism also relies upon the case studies cited by the third argument, but has the logical form of a disjunctive syllogism. Grant the premise that rationalism and the Kuhnian view (broadly understood) are the *only* viable candidates for an account of theory choice. If some of the case studies are incompatible with the Kuhnian view, then it must (tentatively!) be regarded as refuted, implying that the rationalist view is correct. The particular target here is Kuhn’s (1962/1996) holist view on scientific theories or paradigms. This view is in evidence, for instance, when he compares a scientist’s changing of allegiance between paradigms to the “gestalt switches” that occurs when one looks at a duck-rabbit picture. One can ‘see’ either the duck or the rabbit, but not both simultaneously. In the same vein, he describes competing paradigms as “incommensurable” and states that proponents of different paradigms “live in different worlds”. Since holism implies that one cannot simply accept the most important posits of a theory in isolation from the rest of the theoretical structure, this view would therefore appear to be refuted by cases of theoretical synthesis, two of which are discussed in sections 5 and 6.

## 4. The Copernican revolution

Worrall’s (1990) claims the Copernican revolution as an important positive instance for his own rationalist account. He argues that Copernicus’ model of the solar system was both superior to the competing Ptolemaic model by the standard of novel predictive success and eventually victorious because of this superiority. This is in direct opposition to Kuhn, who interprets this case study as a positive example for his own account. The historical and technical details of this section are very largely drawn from Kuhn’s (1957) own detailed study of this episode.

Ptolemy’s *Almagest*, produced around 150 AD in Hellenistic Egypt, proposed a detailed geocentric or geostatic model of the universe. Despite some innovations accumulated during the Middle Ages, this was substantially the system to which Copernicus was introduced in the late 15th century. It envisages a universe contained between two concentric spheres. The surface of the earth is the inner sphere, and the much larger outer sphere holds “fixed stars”. The outer sphere rotates around the inner, with the period of rotation defining a day. In between these two spheres move the planets, a term which applies to the sun, moon, Mercury, Venus, Mars, Jupiter and Saturn. The sun and moon appear as discs, whereas the other planets, like the fixed stars, appear as mere points of light. All the planets move with the outer sphere, but also move slowly across the sphere along a path called the ecliptic. The period of the sun’s path once through the ecliptic defines a year, whereas that of the moon defines a lunar month.

The motion of the remaining planets is more complex, and accounting for their behaviour was the major task of astronomers in both the ancient and early modern periods. Like the sun and moon, the planets progress along the ecliptic, but are not strictly tied to it and may be found as much as 8° away from it on either side. Moreover, while each planet has a characteristic period of movement around the ecliptic, this is only an average; the speed of progression varies over time. In addition to varying in speed, the planets will occasionally reverse course entirely in a so-called “retrograde motion”, before continuing in their usual direction along the ecliptic (see Figure 5). The two “inferior” planets, Mercury and Venus, are each to be found within a limited angular distance from the sun, progressing and then retrogressing back and forth across its position in the sky. The remaining “superior” planets may be found at any angular distance from the sun, but only retrogress when they are on the opposite side of the ecliptic to it (“in opposition”).

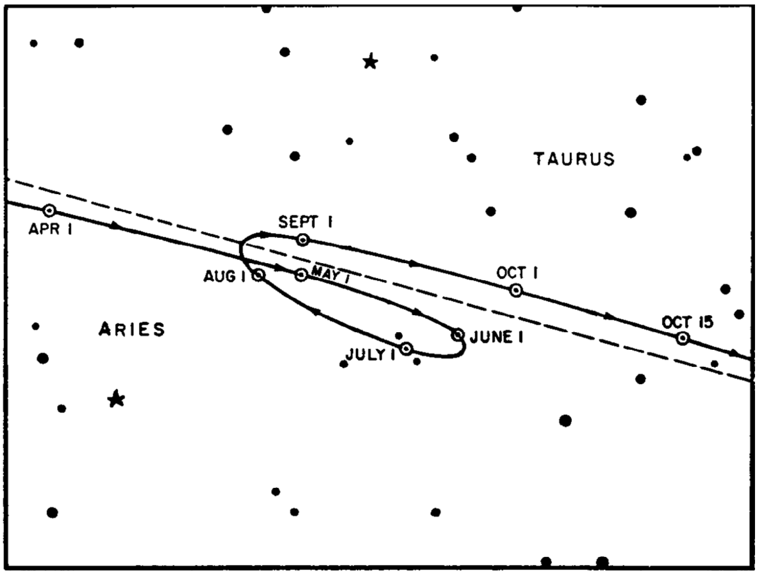

**Figure 5.** Retrograde motion of a planet relative to the fixed stars (reproduced from Kuhn, 1957, p. 48).

Ancient and medieval astronomers devised various mathematical devices to account for these complex observed motions. The focus in this section will be on the best-known of these, namely the deferent-epicycle system. This posits that each planet is carried not just on a great circle centred on the earth (the deferent), but on a smaller circle (the epicycle) carried on the larger circle. The combination of the motion of the two circles produces the motion of the planet. Importantly, the motion of the planet along the epicycle contrary to the overall direction of motion along the deferent gives rise to retrograde motion (see Figure 6). Fixing the speed of rotation of the epicycle fixes the frequency of retrograde motion.

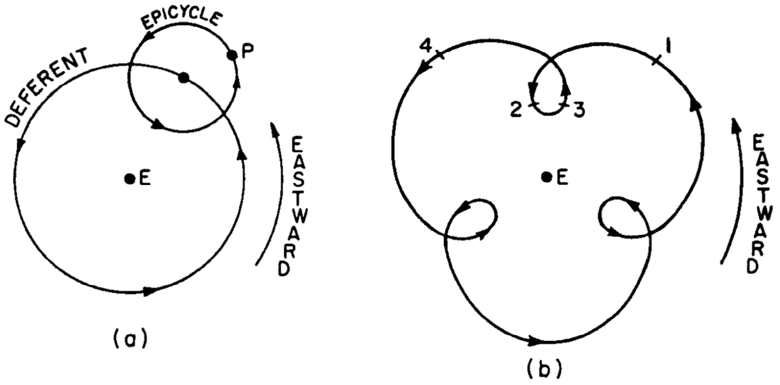

**Figure 6.** The deferent-epicycle system (reproduced from *ibid*, p. 61). Panel (a) shows the overall construction of the system; panel (b) shows the resultant apparent motion of the planet

The Copernican system, introduced with the publication of *De Revolutionibus* in 1543, remained substantially in the Ptolemaic tradition of astronomy in terms of its mathematical techniques and of the basic data set it took as evidence. Its key innovation is that it posited a heliocentric or heliostatic universe, with an immobile sun at the centre, and the earth and other planets revolving in circles around this fixed point. It placed the known planets in the order that is accepted today: Mercury innermost, followed by Venus, then the earth, Mars, Jupiter and finally Saturn. The motion of the moon remained centred on the earth. The key advantage of this system is that it accounts for many of the basic features of the observations very naturally, without the need for epicycles. The distinction between inferior and superior planets emerges naturally – the former are simply within the orbit of the earth, whereas the latter are further out. The retrograde motion of the planets is also explained very simply by the fact that the earth itself is a moving observation platform. It will occasionally 'overtake' the superior planets, or in the case of the inferior ones, be overtaken by them. In both cases the effect is that the planet *appears* to reverse its direction of motion. It also explains why the superior planets should only undergo retrograde motion when in opposition to the sun (the earth passes between the planet and the sun as it overtakes the planet), while the inferior planets pass back and forth regularly across the position of the sun.

However, while Copernicus easily accounts for these gross qualitative features of the data, his system does not yield precise quantitative prediction any more easily than does Ptolemy's. Indeed, Copernicus' system employed just as many epicycles to achieve roughly the same level of accuracy as contemporary geostatic systems did. Moreover, and perhaps more importantly for the reception of the theory, Copernicus' system contradicted several claims that were thought to be established pieces of knowledge at the time. For instance, by removing earth from the centre of the universe, the Copernican system was incompatible with Aristotelian physics. This system of physics, among other postulates, explained motion under gravity by stating that various substances have an intrinsic tendency to move either towards (earth, water) or away from (air, fire) the centre of the universe.

A problem which was especially pertinent to professional astronomers was that Copernicus' model required a vastly larger universe than had hitherto been postulated. The easiest way to see this is to imagine observing a particular star from the earth. As the earth moves in a large circle around the sun, the position of the observer changes and so the observed relative positions of the stars ought to change as well. This phenomenon is known as parallax and is easily observed, for instance, by observing some scene first through one's left eye only and then through the right eye only. One should notice that the relative positions of objects appear to change, and that the effect is more noticeable for objects closer to one's vantage point. That there was no parallax observed for any of the fixed stars at that time (and was not measured at all until 1838) implied either that the earth did not move, or that the sphere of the fixed stars was many thousands of times larger than the orbit of the outermost planet. It may be noted that there was, in Copernicus' time, no direct observational evidence for accepting any particular value of the distance to the fixed stars. Yet the existing cosmology posited that the motion of the entire system originated with the rotation of the fixed stars, which was thence mechanically transmitted to the spheres of the planets. Since such a mechanism would require direct contact between successive spheres, it obviously could not be at work if a vast distance separated the fixed stars from the planets.

Many of the objections to the heliostatic system were met by subsequent astronomers. In 1609, Kepler was able to produce remarkably accurate predictions of planetary positions by positing elliptical orbits. He thus discarded altogether the system of perfect circles and the associated requirement of ever-more-elaborate epicycles. In the same year, Galileo's began making telescopic observations of the heavens. These observations produced powerful new arguments in favour of the heliocentric model. The observation of Jupiter's satellites, for instance, provided concrete evidence of bodies that orbited something other than the earth. And the observation of the phases of Venus showed that this planet could not orbit around the earth. These developments changed the terms of the debate decisively in favour of the heliocentric universe, and scientific opinion came around relatively rapidly. As Kuhn puts it, "By the middle of the seventeenth century it is difficult to find an important astronomer who is not Copernican; by the end of the century it is impossible." (*ibid,* p. 227)

However, it is noteworthy that Kepler and Galileo were convinced by the evidence available for the Copernican model *before* they made their contributions, notwithstanding the difficulties with the theory mentioned above. How are we to assess the rationality of these scientists and their opponents? For Kuhn, the

"Copernican revolution" is the primary illustrative case for his thesis that rationality is underdetermined. There are several theoretical virtues in play. Because their detailed quantitative predictions match the observations about equally well, empirical accuracy does not support either system. Simplicity tends to favour the heliocentric hypothesis, since so many of the broad qualitative features of the data fall out of it naturally. Consistency, on the other hand, favoured the heliocentric model, since it does not require any major modifications to either the physics or the other astronomical beliefs of the time. If the virtues give conflicting judgements, there is nothing intrinsically rational or irrational in preferring one hypothesis over the other. So neither the Copernicans, like Galileo and Kepler, nor their Ptolemaic opponents can be faulted for lacking rational justifications for their positions.

For Worrall (1990), however, once the basic Copernican model had been proposed, accepting it was clearly *more* rational than acceptance of the Ptolemaic model. The latter achieves a fair degree of empirical accuracy, but only at the expense of great theoretical complexity. Virtually all of the interesting empirical consequences of the theory, such as the retrogression of the planets, are used in setting appropriate values for the size and rotational speed of the deferents, epicycles and so on. This model therefore achieves neither use-novel predictive success nor theoretical unification in predicting these consequences. In contrast, a great many of the features of planetary motion that had been observed by that time – such as retrograde motion – simply "drop out" of the basic Copernican model. These particular predictions are therefore use-novel and the overall theory counts as unifying. Moreover, Kuhn's suggestion that the Copernican model sacrifices credibility by contradicting existing scientific theories is forthrightly rejected by Worrall, as noted in section 3.

So, under Worrall's form of rationalism, the Copernican model was clearly rationally superior to the Ptolemaic model. In addition, the Copernican model (with Kepler's modifications) did come to dominate astronomy, so the descriptive claim of rationalism is satisfied by this case. Moreover, as demanded by partial rationalism, the genuine empirical regularities entailed by the Ptolemaic model, such as the occurrence of planetary retrogressions, were at least approximately preserved in the later theory. The partial rationalist who holds that unification is the sole theoretical virtue can therefore claim the Copernican revolution as an example in her favour.

## 5. The neo-Darwinian synthesis

The "neo-Darwinian" synthesis occurred in the early twentieth century and, in broad terms, involved the synthesis of ideas from Mendelian genetics with the Darwinian theory of evolution by natural selection. The historical details of this case are largely drawn from Bowler (2009) and Gould (2002).

Darwin defended two major theses in his 1859 *Origin of Species*. Firstly, he claimed, all contemporary species have evolved from earlier forms, and all forms of life on earth are ultimately derived in this way from a common ancestor. Secondly, he proposed a *mechanism* for this process of evolution, namely natural selection. Natural selection occurs whenever there is heritable variation in phenotypes between organisms, and different phenotypes are associated with different levels of reproductive success. Those characteristics which are associated with reproductive success will gradually increase in frequency in a population, possibly resulting in a population whose features differ markedly from those of the original population, i.e. a new species. In the years following publication of *Origin*, Darwin's first thesis was largely accepted by the scientific community. Indeed, Darwin was not the first to propose it, though he did amass far more evidence for it than any earlier author had achieved.

The second thesis, however, was not so widely accepted. One major reason for this is that a "blended" theory of inheritance was dominant in the latter part of the nineteenth century, and this seemed to be incompatible with the evolution of distinct species by natural selection (Jenkin, 1867). Under a blended inheritance theory, any given trait of an organism will have a value that is intermediate between that of the corresponding values in its parents. The child of a tall mother and a short father, for instance, will typically be of middling height. However, any individual organism that differs markedly from the remainder of the population must eventually breed with more typical individuals. Thus, even if the individual's unusual phenotype represents an adaptive advantage, none of its descendants will possess this phenotype to the same degree. Thus it will be extremely difficult for a novel phenotype to spread through the population, and so for speciation to occur.

The blended theory of inheritance was eventually overthrown by a "particulate" theory of inheritance, which holds that many or most phenotypic features are propagated from parent to offspring in discrete, indivisible heritable elements (now referred to as "genes"). A trait can therefore, under many circumstances, be passed completely intact from parent to offspring. Indeed, even if the phenotype associated with a given gene fails to be expressed in immediate offspring due to the effects of other genes, the gene can still circulate in the population and be expressed in later generations. The particulate theory of inheritance is derived from a very natural interpretation of Mendel's 1865 paper *Experiments on Plant Hybridisation*, although he does not explicitly propose it himself. The theory only began to be disseminated when Mendel's paper was "rediscovered" by the broader community of biologists from 1900 onwards.

Mendel's discoveries were initially interpreted as a further blow to the thesis that speciation occurs *via* natural selection. Particulate inheritance would seem to imply that differences between organisms must be relatively large and discrete. So new species should typically appear *via* "jumps" or saltations, not *via* the gradual modification of existing traits. Advocates of saltationism included de Vries, Correns and Bateson. They were opposed by the so-called "biometricians", including Weldon and Pearson, who emphasised the continuous variation of phenotypes within a population. Both parties had some empirical evidence in their favour. Indeed, in his 1894 book *Materials for the study of variation*, Bateson provided many actual examples of both of characters that vary continuously in a population and of characters where there is discontinuity. Neither party, however, was able to provide a fully satisfactory theory of evolution.

Fisher (1918; 1930) made a breakthrough in resolving the biometrician-saltationist dispute by arguing that a given phenotypic feature may be governed by the cumulative effect of several distinct Mendelian genes. This renders a particulate mechanism of inheritance compatible with apparently continuous phenotypic variation for many traits. In addition to dissolving the saltationist-biometrician dispute, Fisher's insight provided a plausible mechanism for speciation by natural selection. If a trait is favoured by natural selection under certain circumstances then, all else being equal, genes associated with that trait will increase in frequency within a given population. The degree to which that trait is expressed in individual organisms possessing the gene need not be reduced in successive generations. Moreover, if a particular subpopulation is subject to selective pressures which the majority of the population is not, it may accumulate genetic differences to the extent that it comes to form a distinct species. Fisher's approach, under the label of "population genetics" rapidly came to dominate theoretical biology.

The brief account given above neglects to mention many interesting features of the neo-Darwinian synthesis, and many scientists who were important in its development. Nevertheless, it should be clear that it is an instance of theoretical synthesis as described in section 3. Both the Darwinian/biometric research programme and the Mendelian/saltationist programme were well-supported by empirical evidence. Nevertheless, they were initially mutually incompatible. The crisis was resolved because scientists such as Fisher were able to identify the elements of each theory essential to its predictive success, and weaken their commitment to those elements that are not essential. For instance, the standard explanation of Mendel's results is based on two assumptions: (a) that inheritance is particulate; and (b) that the traits under observation are governed by the activity of only one, or a small number of, genes. The modern synthesis is partially based upon recognising that the second assumption need not apply to organisms generally[18]. And, once this assumption is understood as a contingent boundary condition rather than an essential element of the "Mendelian" theory, the conflict

[18] Indeed, Mendel needed to breed his plants for some time to obtain "pure" strains that exhibited the observed pattern of inheritance.

between it and the Darwinian theory is resolved, yielding a synthesis which was more empirically successful than either of its constituents.

## 6. The prion revolution

The “prion revolution”, which largely took place between the 1960s and the 1990s, is another example of a major theoretical change in biology. The historical details of this episode are taken from an excellent pair of papers by Keyes (1999a, 1999b). For the sake of brevity, the following exposition will be admittedly somewhat “Whiggish”, focusing largely on hypotheses and experimental results which, in hindsight, led towards the theoretical approach which is now accepted.

A key part of the background to this episode is the so-called “central dogma” of molecular biology (see Crick, 1958; 1970). The paradigmatic set of processes associated with the dogma begins with the transcription of a portion of DNA to a complementary strand of messenger RNA. This is then translated into a corresponding protein. The eventual sequence of amino acids in the protein is encoded in the sequence of bases in the DNA. Sequence information can pass from RNA into DNA by the action of reverse transcriptase enzymes. However, “once ‘information’ has passed into protein *it cannot get out again”* (Crick, 1958), p. 153, emphasis in original*).* The dogma asserts, in other words, that there is a *one-way* flow of information from nucleic acids to protein sequence. An important corollary of this assertion is that any biological agent which replicates must contain a nucleic acid, as replication requires the propagation of sequence information.

A major challenge to the dogma arose indirectly, through the investigation of a disease called scrapie. This disease affects primarily sheep but can also infect other mammals, and is characterised by severe and invariably fatal degeneration of the brain tissue. It has been known since the 1930s that the disease can be transmitted to healthy animals by inoculating them with infected tissue. In the 1960s, however, it started to become apparent that the scrapie pathogen did not resemble any other known infectious agent. Firstly, scrapie can be transmitted by a suspension of brain tissue that has been filtered so finely that no whole cells (including bacterial cells) could be present. This result is compatible with the infectious agent being a virus. However, infected tissue was also found to be infectious after treatment with heat or chemical agents which would normally inactivate any viruses present. Infectivity was also preserved after treatment with ultraviolet radiation of a wavelength particularly destructive to nucleic acids. Further studies of the infectious agent revealed that it is far smaller than any known virus, and is more comparable in size to protein molecules. Such a small particle is unlikely to harbour a nucleic acid molecule long enough to code for even a single protein. These results together suggested that the infectious agent does not in fact rely upon nucleic acids for its replication at all.

These results sparked a wave of speculation as to the chemical identity of the scrapie agent. None of these speculations could be confirmed or refuted, however, until the scrapie agent was purified and examined in more detail. Much of this work was carried out at by Prusiner and colleagues at the University of California in San Francisco (UCSF). In the early 1980s, this team (Prusiner *et al.*, 1981; Prusiner, 1982) demonstrated that the scrapie agent was inactivated by substances and processes known to disrupt protein structure, but which do not affect nucleic acids, and was conversely unaffected by enzymes which destroy nucleic acids. Together, these results suggested that the agent is a protein, or “prion” (for *pro*teinaceous *in*fectious particle). The UCSF group subsequently isolated a particular protein which they concluded is the scrapie agent and which they therefore named PrP (for “prion protein”, (McKinley *et al.*, 1983; Prusiner *et al.*, 1982). This conclusion was supported by several experiments. For instance, whenever PrP was obtained from brain tissue, even by different purification methods, the resulting fraction was infectious. And chemical treatments which denatured PrP also eliminated the infectivity of a sample.

Pattison and collaborators had, in the 1960s and 70s demonstrated that scrapie could sometimes occur spontaneously in animals, or after ‘infection’ with the brain tissue of a healthy animal. To Pattison (1982), this suggested that infection with scrapie was simply the “unmasking” of a entity already present in the brain. This hypothesis would also explain the inability of the UCSF group to find nucleic acid in infectious particles. To test Pattison’s hypothesis, part of PrP protein was sequenced, and a probe was constructed to match the corresponding DNA sequence (Oesch *et al.*, 1985). Using this probe, it was found that the relevant gene is, indeed, found in the genome of healthy animals. Moreover, the corresponding messenger RNA is found in both healthy and scrapie infected animals, and at identical concentrations. This suggests that the PrP protein is produced in healthy animals at the same rate that it is in infected animals. Indeed, it was soon found that a chemically distinct version of this protein (labelled $\text{PrP}^{\text{C}}$) was present in both healthy and infected animals. The apparent upshot of these results is that normal $\text{PrP}^{\text{C}}$ is produced in the brain of all scrapie-susceptible mammals and, in infected animals, is somehow converted into the disease-associated form $\text{PrP}^{\text{Sc}}$.

These results were compatible with a “protein-only” hypothesis of scrapie transmission. This hypothesis continued to face resistance, however, on the grounds that it contradicted the extremely well-confirmed central dogma (see, for instance, (Carp *et al.*, 1985; Gajdusek, 1985). Moreover, the prion theorists had not yet provided evidence for any plausible mechanism by which a protein might carry “genetic information”. One possible mechanism, suggested by Griffith (1967), is that the protein might induce its own expression in host cells. This hypothesis, however, is not compatible with the fact that healthy and infected tissues apparently produce PrP at the same rate. Moreover, from the 1970s it became apparent that there are several distinct “strains” of scrapie. This, together with the fact that there is only one copy of the PrP gene on the mammalian genome, would seem to suggest that infectious particles do not merely unlock some potential that is latent in cells, but carry strain-specific information which is replicated in daughter particles. The emphasis therefore shifted to another version of the hypothesis, that the scrapie protein $\text{PrP}^{\text{Sc}}$ *directly* converts the normal $\text{PrP}^{\text{C}}$ into copies of itself. Versions of this hypothesis were discussed by Griffith, by the UCSF team (Oesch, *et al.*, 1988; Prusiner, 1984), and by Bolton and Bendheim (1988).

Keyes (1999b) emphasises that this sort of hypothesis by itself represents a significant departure from the central dogma, because it distinguishes between *types* of biological information. *Sequence* information is still transmitted in the standard way from nucleic acids to proteins. However, information might also be encoded in the post-translational state of the protein and propagated 'horizontally' between protein molecules. There were, however, by the late 1980s still several viable candidates for exactly what form this encoding might take. Protein molecules can be chemically altered in various ways, by the modification of an amino acid residue, or by the addition of a chemical group. However, lacking any evidence of chemical alteration, Prusiner and co-workers (1990) favoured the hypothesis that the difference between $PrP^{C}$ and $PrP^{Sc}$ is *conformational*. "Conformation", in this context, refers to the exact three-dimensional shape that a two-dimensional chain of amino acids adopts after translation. Although conformation is of course dependent on the linear sequence, it is not at all rigidly determined by it[19]. Importantly, the conformational hypothesis is able to explain the existence of different strains of scrapie, as there can be many conformational variants of a given protein.

This conformational hypothesis has been steadily accepted by the scientific community since 1990, with further supporting evidence emerging from various sources. Prusiner himself won the Nobel Prize in 1997, indicating that the theory was understood to represent well-established knowledge. Moreover, it is now standard to find some discussion of prions in the chapter on protein folding in mainstream biochemistry textbooks (see, for instance, Nelson & Cox, 2005, ch. 4). There are, of course, scientists who continue to reject all versions of the protein-only hypothesis and promote nucleic acid theories of scrapie, although these represent a dwindling minority.

Even allowing for the possibility that the prion theory is mistaken, it is worth outlining how the present consensus represents a *synthesis* of two apparently

[19] Indeed, at the time or writing, it is still beyond our capabilities to predict the final stable conformation of a protein given its amino acid sequence. This is the so-called "folding problem".

incompatible theories. The central dogma of molecular biology states that biological information can only be replicated in the form of nucleic acids. The prion theory of scrapie states that biological information can be replicated ‘horizontally’ between proteins. The current synthesis between the two is achieved by distinguishing between *sequence* and *conformational* information, then narrowing the domain of the central dogma to include only the former. Thus all the known empirical successes of this dogma are accounted for, while also accommodating the empirical success of the prion theory.

## 7. Chapter summary

In this chapter, the problem of rationality in theory choice in science was introduced *via* Kuhn’s treatment of this topic in *Structure*. Writing in 1962, Kuhn appears to imply that instances of theory choice are not at all governed by objective standards of rationality. This radical interpretation of Kuhn’s early work is endorsed by proponents of the Strong Programme, who deny any role to objective standards of rationality. They claim that which of several competing theories ends up dominating the field is determined largely by the political and/or social power of these theories’ respective supporters. The deficiency of the Strong Programme is that it is *too* radical – objective (or at least widely shared) rational standards surely play at least *some* role in theory choice.

In this chapter, three views which accord some role to rational standards have been surveyed. The first is so-called “Whig history”, which designates as rational just those historical theories which can be viewed as predecessors of our current best theories. This was rejected on the grounds that it does not really provide any standards of rationality which can sensibly be applied prospectively against a background of already-accepted beliefs. The second is Kuhn’s later view, which holds that there are objective standards of scientific rationality, but that these underdetermine theory choice in many or most actual cases. Scientific decisions are therefore at least partly dependent on non-rational factors, such as the aesthetic preferences of individual scientists. The third view is rationalism, which holds that it is usually possible to determine which of several competing theories it is *more* rational to believe and moreover that the most rational preferable theory has indeed come to prominence in most actual instances of theory change.

In this chapter, a particular version of rationalism called “partial rationalism” has been defended. Partial rationalism claims that scientists are rational just in case they accept those *parts* of theories responsible for empirical successes (and have largely acted rationally by this definition). This view has been illustrated (although, of course, not generally proven) by several case studies. Following Worrall, it has been demonstrated that the victory of the Copernican over the Ptolemaic model of the solar system represents the victory of the more empirically successful theory (by either the use-novelty or the unification standard). In the case of both the neo-Darwinian synthesis and the prion revolution, it has been demonstrated that the theory which was eventually accepted is constituted by (among other elements) the essential parts of the ostensibly competing theories.

The following chapter will focus upon some case studies which are, at least *prima facie*, problematic for partial realism, and thus for partial rationalism. After examining these cases in detail, some remarks will therefore be made on the tenability of partial rationalism.

# Chapter 5. Problem cases

## 1. Chapter overview

In this chapter, several putative counterexamples to deflationary realism, and thus to partial realism and partial rationalism, are examined in detail. The stage is set in section 2, where several lists of potential counterexamples that have been proposed by other authors are reproduced. From these lists, two particular cases are selected for more detailed study, namely the miasma theory of disease and the phlogiston theory of chemistry. They are chosen because they are both cases of theories that were, at least on the face of it, empirically successful, and yet their essential posits were not preserved in successor theories.

In section 3, the miasma theory of disease is examined particularly in the context of nineteenth century England. Several notable developments occurred in this period. Firstly, there were large scale changes to urban water supply and sewage infrastructure, and these changes were partly motivated by the miasma theory. Secondly, England, and particular London, was struck by epidemic cholera, the transmission of which was at least initially blamed on atmospheric miasmas. Thirdly, the miasma theory, at least insofar as is applies to cholera, was eventually rejected in favour of a waterborne theory of transmission. In section 4, it is demonstrated that, although changes in public health infrastructure did eventually eliminate the spread of cholera, these changes were ineffective and in fact counterproductive when carried out under the influence of the miasma theory. It was only the waterborne theory of transmission that allowed for effective intervention in the spread of disease. Thus the miasma theory was (at least in this episode) not empirically successful and so fails as a counterexample to DR.

Section 5 gives a briefly outline of the evolution of chemical theory over the course of the eighteenth century, during which the phlogiston theory gave way to Lavoisier’s oxygen theory. In light of this history, in section 6, it is argued that the phlogiston case presents a significant challenge to the scientific realist. This theory *was* empirically successful, and applies concepts that are structurally similar to those found in modern electrochemistry. These concepts are not, however, preserved in the immediate successor theory to phlogiston, namely the oxygen theory. It is argued that this course of events poses a serious challenge to the descriptive component of the realist view, i.e. the claim that the essential elements of successful theories will generally be preserved in instances of theory change. Thus deflationary realism and the descriptive component of partial rationalism are also challenged. The normative components of these views, and the claim that such elements accurately represent the world, are not, however, challenged.

## **2. Putative counterexamples to realism**

In this thesis so far, several historical case studies have been discussed for the purposes of illustrating realist and/or rationalist views. However, having thoroughly described these views in the preceding chapters, it is time to examine some putative counterexamples to them. As stated in Chapter 1, a putative counterexample to deflationary (and thus partial) realism has the form of a historical episode in which some theory is empirically successful, certain elements of the theory are essential to this success, and yet this element is not preserved in successor theories. Several authors have offered lists of historical episodes which purportedly match this template. These will be reviewed briefly below. Following this, two of them will be examined in more detail, to determine whether they do, in fact, stand as counterexamples.

Laudan, in support of his pessimistic meta-induction, offers a list of theories "which were both successful and (so far as we can judge) non-referential with respect to many of their central explanatory concepts" (Laudan, 1981, p. 33):

- the crystalline spheres of ancient and medieval astronomy;
- the humoral theory of medicine;
- the effluvial theory of static electricity;

- 'catastrophist' geology, with its commitment to a universal (Noachian) deluge;
- the phlogiston theory of chemistry;
- the caloric theory of heat;
- the vibratory theory of heat;
- the vital force theories of physiology;
- the electromagnetic aether;
- the optical aether;
- the theory of circular inertia;
- theories of spontaneous generation.

Several of these examples have already been discussed in detail by scientific realists, with the aim of discrediting them as counterexamples. Frequently, the approach has been to argue that the particular theoretical posit highlighted by Laudan was *not*, in fact, essential to the empirical success of the theory in question. Following several papers by Worrall (1989a, 1989b), for instance, it is a standard realist tactic to argue that the posited existence of an “optical aether” did not play an essential role in the successful predictions of nineteenth-century theories of light. Psillos (1999) has offered a similar analysis for the posited existence of caloric. Another frequent tactic is to argue that the theory highlighted by Laudan was not sufficient empirically successful to warrant application of the no-miracles argument in the first place. For instance, although Worrall (1990) does not discuss “crystalline spheres” per se, he does adopt this approach in his treatment of the Ptolemaic model of the solar system (discussed in Chapter 4.4).

Stanford (2006) shifts emphasis away from Laudan’s “pessimistic induction, preferring instead to talk about the “problem of unconceived alternatives”. He argues that scientists throughout history have simply failed to conceive of theories that were at least equally well-confirmed by the evidence available to them as the theories they actually did accept. We know that such alternatives existed because our *modern theories* are, of course, well-confirmed by this evidence. He therefore proposes a “new induction” focusing on the epistemic situation of scientists themselves, with the conclusion that we are likely to be in the same position as these historical scientists. That is, we are unable to imagine the theories that would account for our empirical observations (and presumably also deal with anomalies) better than our current theories do. Stanford's argument is clearly, as Lewens (2006) puts it, a "close relative" of the pessimistic induction. Whatever the exact nature of this relationship, the problem of unconceived alternatives, like the pessimistic induction, is effective as an argument against partial realism insofar as it can point to examples of where the "essential elements" of empirically successful theories were not preserved in successor theories. Stanford's examples, to a certain extent overlapping with Laudan's, are as follows:

- from elemental to early corpuscularian chemistry to Stahl's phlogiston theory to Lavoisier's oxygen chemistry to Daltonian atomic and contemporary chemistry
- from various versions of preformationism to epigenetic theories of embryology
- form the caloric theory of heat to later and ultimately contemporary thermodynamic theories
- from effluvial theories of electricity and magnetism to theories of the electromagnetic ether and contemporary electromagnetism
- from humoral imbalance to miasmatic to contagion and ultimately germ theories of disease
- from eighteenth century corpuscular theories of light to nineteenth century wave theories to the contemporary quantum mechanical conception
- from Darwin's pangenesis theory of inheritance to Weismann's germ-plasm theory to Mendelian and then contemporary molecular genetics
- from Cuvier's theory of functionally integrated and necessarily static biological species and from Lamarck's autogenesis to Darwin's evolutionary theory.

Another author who has offered such a similar list is Lyons (2002). He has, however, attempted to respond to realist criticism of Laudan and Stanford's lists by citing only episodes in which the theory concerned achieved novel predictive success (by the use-novelty account). Nevertheless, his list still overlaps somewhat with their lists. It is as follows:

- Caloric theory
- Phlogiston theory
- Rankine’s nineteenth century vortex theory (of thermodynamics)
- Newtonian mechanics
- Fermat’s principle of least time (in optics)
- Maxwell’s ether theory
- Dalton’s atomic theory
- Kekulé’s theory of the benzene molecule
- Mendeleev’s periodic law
- Bohr’s 1913 theory of the atom
- Dirac’s relativistic wave equation
- The original (pre-inflationary) big bang theory

Although, from the partial realist standpoint, Lyons is to be credited for including only predictively successful theories, it is highly contentious that in all these cases the relevant essential elements were not preserved in successor theories. As argued in Chapter 1.6, both Newtonian mechanics and Mendeleev’s periodic law, for instance, are at least approximately preserved.

Finally, Vickers (forthcoming) lists no fewer than twenty-one putative cases of predictively successful theories that are not even approximately and/or partially true from the perspective of later theories. Since he explicitly draws on the lists given by Laudan, Lyons and Stanford, only items not suggested by these other authors will be cited here. These include:

- The teleomechanist theory of biological development
- The Titius-Bode law of planetary orbits
- Kepler’s prediction of the rotation of the sun

- Kirchhoff’s theory of diffraction
- Sommerfeld’s prediction of the hydrogen fine structure
- Velikovsky and Venus (as discussed in Chapter 2.10)
- Steady-state cosmology
- The achromatic telescope (as discussed in Chapter 3.8)
- The momentum of light
- S-matrix theory of particle physics
- Lorentz’s theory of variation in electron mass with velocity
- Taking the thermodynamic limit

All of the cases cited by the authors above deserve further philosophical consideration. However, in the remaining sections of this chapter, two in particular will be examined in further detail. These are the miasma theory of disease, which is included on Stanford’s list; and the phlogiston theory in chemistry, included by all of the authors cited above.

These two cases are chosen because a strong *prima facie* case can be made for each of them as a counterexample to deflationary realism (DR), as it has been defined in Chapters 1-3 of this thesis. Recall that DR is primarily committed to the continuity of science. As such, it is a logical consequence of partial realism and is identical to the descriptive component of partial rationalism. Thus, if either of the case studies carried out below refutes DR, it also refutes these more substantive views. However, at certain points, the implications of a particular historical turn of events for partial rationalism will be addressed directly. This is because the idea of theoretical synthesis, although it also follows as a logical consequence of DR and PR, has been addressed primarily with reference to this position.

## 3. The miasma theory of disease and the contagionist revolution

The miasma theory of disease maintains that the primary cause of at least certain diseases is atmospheric pollution arising from rotting organic matter. The first recorded mention of this theory is attributed to one of the great authors of classical antiquity, Hippocrates (1978), who makes the connection between marshy terrain and fevers. This connection is still preserved in the name of the disease malaria, literally “bad air” in Italian (Reiter, 2000). This theory poses a *prima facie* challenge to DR because it posits a certain mechanism for the spread of disease – namely, atmospheric pollution – which modern scientists universally agree does not exist. Moreover, as discussed below, it appears that this theory was at least somewhat empirically successful. Although this theory enjoyed some currency among medical practitioners for at least two millennia, the focus on this chapter will be nineteenth century Britain, where it was eventually rejected by the medical community.

The major competitor to the miasma theory by the early nineteenth century was the contagion theory of disease. This holds that disease is primarily transmitted from afflicted to healthy individuals by close contact. The contagion theory, in fact, has origins that are just as ancient as those of the miasma theory. These are articulated indirectly, for instance, in various prohibitions on contact with “unclean” persons or substances found in the Jewish Old Testament (Ackerknecht, 1948/2009, p. 8). The contagion and miasma theories in fact co-existed, although somewhat uneasily, during the early nineteenth century. Ackerknecht characterises the medical profession at this time as comprising two “extreme wings” of committed contagionists and miasmists, with a moderate and ecumenical centre of “contingent contagionists” – those who believed that contagion was possible, but only under certain circumstances. Moreover, even the most extreme miasma theorists acknowledged certain cases of contagion – syphilis, for instance, was universally understood to spread directly *via* sexual contact. So, in practice and despite their ideological commitments, medical professionals typically applied theoretical frameworks to particular diseases on a case-by-case basis.

Major changes in theories of disease were catalysed by the series of cholera epidemics that struck Europe in the nineteenth century, beginning in the 1830s. The historical details of these epidemics are largely drawn from  Longmate (1966). Cholera is a disease characterised by extreme intestinal dysfunction, resulting in large quantities of watery diarrhoea and vomiting. It originated in and is endemic to the area around the lower Ganges river and first became a global pandemic in 1817, following the extensive international trade and communication routes that had been established by European colonial empires by that time. This first pandemic, continuing until 1826, did not reach Europe. However, subsequent pandemics affected Britain in the years 1831-1832, 1848-1850, 1853-1854, and finally in 1866. In what follows, the focus will be on the metropolis of London, where cholera epidemics were especially severe, and where many of the important theoreticians carried out their work.

Both the miasma and contagion theories appealed to features of water supply and sanitation in their explanations of the aetiology of cholera, although in different ways. Therefore, before discussing the development of medical theory directly, it is worth providing an overall picture of the evolution in urban water infrastructure that was also in motion at this time. This evolution was characterised by two interconnected developments, namely the abandonment of the cesspool system in favour of flush toilets; and the abandonment of water obtained directly by households in favour of water piped from distant sources. This topic is addressed in illuminating detail by Hardy (1984; 1991) and Halliday (1999).

The ancient system of human waste disposal in cities involved the emptying of chamber pots into cesspools, which were typically shared between a small group of houses. When a cesspool was full, its contents were hauled away by cart, typically to be used as agricultural fertiliser. This elegant system for recycling organic matter started to break down in London as the city expanded, and the cost of hauling material out in the countryside increase proportionately. Even more devastating, from about 1847, vast amounts of cheap guano (bird droppings) became available on world markets, drastically undercutting the price of human manure. In consequence, hauliers charged higher prices for emptying cesspools, and poorer residents were less willing or able to pay for removal. The result is that cesspools were dug deeper to accommodate more waste, contaminating groundwater, or simply allowed to overflow.

Over the first half of the nineteenth century, just as the cesspool system was becoming uneconomical and beginning to pose an increasing threat to human health, flush toilets became steadily more common. The large amounts of waste water produced by a flush toilet must, of course, go somewhere, and so private houses were connected to sewer systems. London had a pre-existing sewage system, but this was intended primarily to channel water from storm drains. This changed in 1815, when it became legal to dump household waste into the sewage system. For reasons that are obvious to the modern reader, it is extremely unfortunate that this sewage system initially emptied directly into the river Thames. Contemporaries, too, became progressively more dismayed at the polluted state of this river, which culminated in the “Great Stink” during the summer of 1858. The Great Stink served as the impetus for the massive programme of sewer construction under Bazalgette, which took place largely in the 1860s. The smaller sewers were consolidated into larger pipes and, although the resulting effluent was still dumped into the Thames, the outlet was far downstream of London. Thus, the extreme pollution of the river was a phenomenon predominantly of the middle part of nineteenth century.

Prior to the nineteenth century, most of the water consumed in London was either obtained from relatively shallow wells, or directly from rivers and canals. However, over the course of the eighteenth and early nineteenth centuries, various private water companies were established to supply piped water to homes and other establishments (Dickinson, 1954). Notice that a copious supply of piped water is also, of course, essential to the functioning of the flush toilets which were being installed at this time. However, because piped water supplies were often beyond the means of poorer residents, and were generally unreliable in any case, many inhabitants of the metropolis continued to depend on groundwater obtained from pumps or on surface water.

Interestingly, many of the changes in water and waste removal infrastructure were motivated directly by the miasma theory of disease. Thus, this theory can be considered empirically successful to the extent that these changes were successful in reducing the burden of disease. The influence of the miasma theory is clear in the writings of the so-called sanitarians who promoted these changes. For instance, Chadwick, a lawyer and Benthamite social reformer, in 1842 authored *The Sanitary Condition of the Labouring Population.* This argued that:

> “the various forms of epidemic, endemic, and other disease [are] caused, or aggravated, or propagated chiefly amongst the labouring classes by atmospheric impurities produced by decomposing animal and vegetable substances, by damp and filth, and close and overcrowded dwellings ... where those circumstances are removed by drainage, proper cleansing, better ventilation, and other means of diminishing *atmospheric impurity*, the frequency and intensity of such disease is abated”
>
> “The primary and most important measures [for improving sanitary conditions] ... are drainage, the removal of all refuse of habitations, streets, and roads, and the improvement of the supplies of water. [The] expense [of removing decomposing refuse] may be reduced ... or rendered inconsiderable, by the use of water and self-acting means of removal by improved and cheaper sewers and drains.” (Chadwick, 1842, section 9, emphasis added)

Notice that Chadwick advocates both flush toilets (“self-acting means” of waste removal) and the improvement of water supplies. It is clear from other remarks, however, that he is less concerned with the quality of *drinking* water than that there was sufficient water that houses may be cleaned and waste washed away more quickly.

Chadwick’s recommendations proved to be hugely influential on government policy (Halliday, 1999, pp. 48-57; Porter, 1999, pp 411-412). The Public Health Act of 1848 established a General Board of Health (of which Chadwick was a member) with the function of enforcing public health standards across England and Wales.

The Board was empowered to force local government entities to establish and maintain sewer systems, clear up refuse, and so on. The Metropolitan Sewers Act of 1848 forcibly amalgamated most of the sewer commissions operating in the London area into a single Metropolitan Sewers Commission (with Chadwick, again, one of the commissioners). The Act required, among other things, that all new houses be built with toilets and drains up to a certain specification, and empowered the Commissioners to order existing houses be connected to the sewage system. The Metropolis Water Act of 1852 required that the water companies operating in London draw their water from a section upstream of the heavily-polluted tidal section of the Thames. The passing of this Act should not, incidentally be taken as an indication that the government believed that disease could be transmitted directly through contaminated drinking water; miasma theorists had their own reasons for objecting to "foul-smelling", polluted water. Moreover, this measure was not treated as a priority. This is underlined by the fact that the Act was very poorly enforced, with several companies failing to comply for more than a decade.

Returning to questions of medical theory, it becomes quickly apparent that Chadwick's acceptance of the miasma theory was not merely dogmatic. Several pieces of evidence appeared to support it quite decisively over the contagion theory. Firstly, various quarantine measures that were enacted in an attempt to limit the spread of the 18331-32 pandemic proved utterly ineffective. Secondly, people and areas who were in regular contact with afflicted persons or regions, without the benefit of quarantine, often did not themselves become afflicted (Ackerknecht, 1948/2009, p. 12). Moreover, in addition to negative evidence against the contagion theory, there appeared to be quite good evidence supporting atmospheric transmission. In what remains a seminal work in the foundation of epidemiology, Farr (1852), a statistician and civil servant working at the General Register Office, produced a report on cholera mortality in England from 1848 to 1849. In this report, he noted an extremely regular connection between the elevation of a London inhabitant's dwelling above the river Thames and his or her chance of dying from cholera. Some of his data are reproduced in Table 1.

| Mean Elevation of the ground above the Highwater Mark (feet) | Mean mortality from Cholera (per 10000) |
|---|---|
| 0 | 177 |
| 10 | 102 |
| 30 | 65 |
| 50 | 34 |
| 70 | 27 |
| 90 | 22 |
| 100 | 17 |
| 350 | 7 |

**Table 1**. The correlation between elevation and cholera mortality in London during the epidemic of 1848-1849 (*ibid*, p. 63)

This was viewed as confirmatory evidence for the miasma theory precisely because, as indicated above, low-lying areas nearer to the Thames were so foul-smelling by this time.

Snow (1849), in his *On the mode of communication of cholera*, proposed a theory that was opposed to both the miasma view and to versions of the contagionist theory which emphasised direct transmission between persons. He suggested that cholera was caused by some “material” which acted to irritate the intestinal wall, and could be spread to others when they inadvertently ingested the discharge of an afflicted individual[20]. Moreover, he argued, the most common means for the spread of this material is *via* the contamination of drinking water. Snow supported this hypothesis by citing various incidents where a cluster of cholera cases had occurred in association with faecal contamination of the water supply. For instance, he describes a group of houses named Albion Terrace in Wandsworth, in which many inhabitants became ill, although none of the other residents of the immediate

[20] Note that Snow wrote before the germ theory of disease became widely accepted, and so he did not think of this “material” as a living organism.

area did. The only thing that obviously distinguished these residents from their neighbours is that they shared a water supply, and that this had become contaminated by the cesspools behind their houses.

Snow produced a substantially expanded edition of his book in 1855, in which he adduced several more lines of evidence for his theory. The most famous of these from the modern perspective is his study of the epidemic of 1854 at Golden Square in Soho. Using records obtained from Farr’s office, he was able to systematically chart the place of residence of cholera victims in relation to the various water pumps in that neighbourhood. A detail of his map is reproduced in Figure 7. What this map showed is that cases of cholera were centred on a water pump on Broad Street, with the number of deaths decreasing with distance from the pump. Moreover, the residents of a nearby workhouse (marked on the map) with its own water supply were relatively unaffected.  In addition, Snow systematically interviewed acquaintances of all known victims, demonstrating that each had at least very probably drunk water from this pump. Perhaps the most impressive such case was that of an afflicted woman who lived in Hampstead, several miles away from Soho. It emerged that she had lived near the Broad Street pump some years before, and had acquired a taste for its water. She therefore had specifically obtained some for her own consumption, and became ill the following day. Finally, Snow was able to demonstrate that the water supply of this particular pump was contaminated by a nearby cesspool. Using this evidence, Snow convinced the authorities of the local parish to remove the handle of the pump (meaning water could not be drawn from it), and the epidemic ceased shortly thereafter. As Snow himself admitted, however, this in itself was not decisive evidence for his theory, as the epidemic was in any case abating by this point.

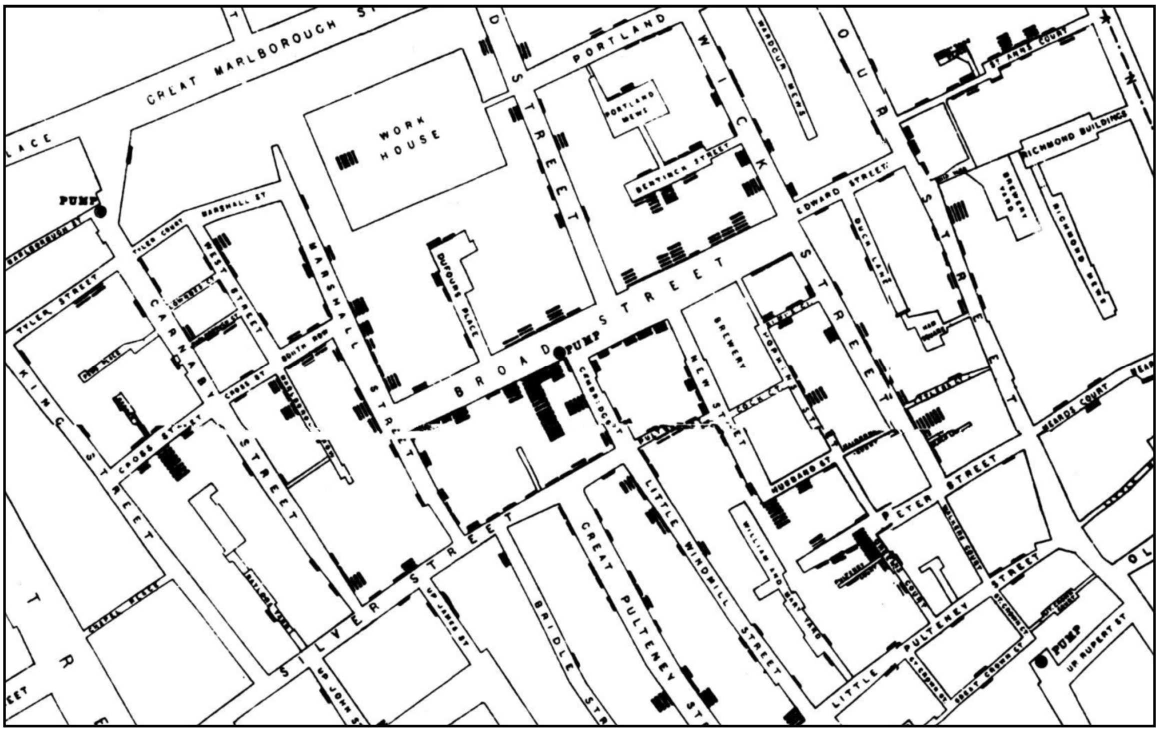

**Figure 7.** A reproduction of Snow’s (1855) map showing the spatial distribution of cholera deaths around Golden Square in Soho. The dashes tally deaths at a given location, and pumps are represented by circular dots. The Broad Street pump is located roughly in the centre of the image.

As pointed out by several authors (see, for example, McLeod, 2000), another analysis of the 1854 epidemic carried out by Snow  was in fact more convincing to his contemporaries. South London, at that time, was served by both the Lambeth and the Southwark and Vauxhall water companies. Moreover, there were certain areas in which they competed for business, even serving different houses on the same street. The geographic distribution of their pipes is depicted in

Figure 8. The Lambeth Company, as it happened, had complied with the Metropolitan Water Act of 1852 and so had moved its water inlet upstream of the tidal part of the Thames. The Southwark and Vauxhall Company, in contrast, still drew from a heavily polluted section of the river. Despite many difficulties in ascertaining exactly which household was supplied by which company, Snow was able to establish to that customers of the Lambeth Company suffered significantly

less during the 1854 epidemic than did those of the Southwark and Vauxhall Company.

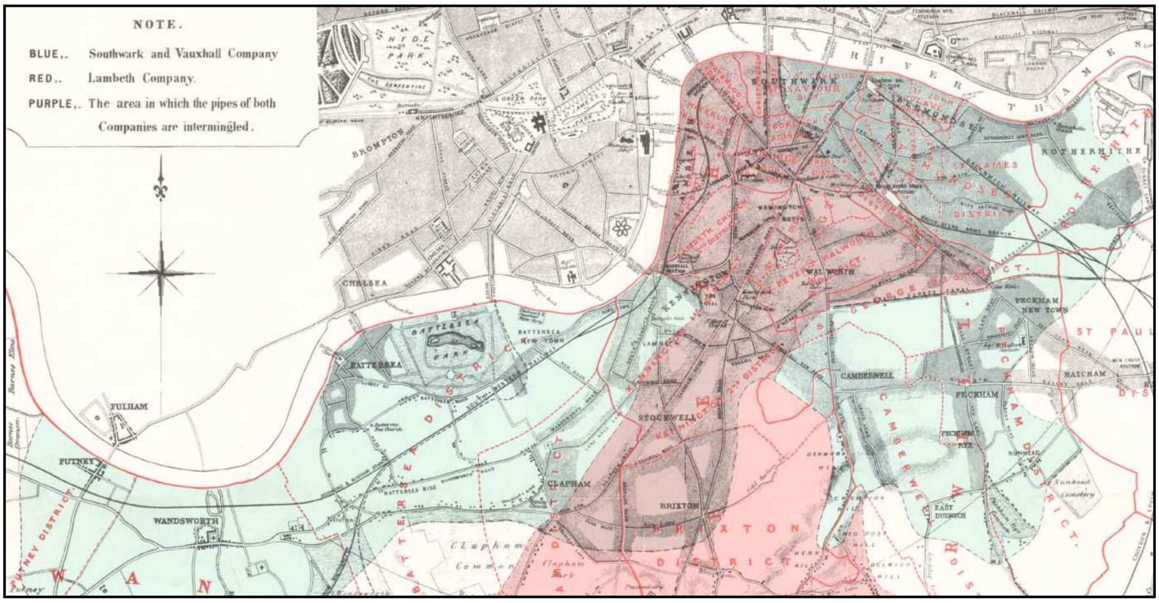

**Figure 8.** A reproduction of Snow’s (1855) map showing the areas of south London to which the Lambeth and the Southwark and Vauxhall companies provided water.

Snow died in 1858, but his analysis of the south London epidemic had convinced Farr of the waterborne theory of cholera transmission, and he became its primary champion (Eyler, 1973; 2001). When mortality reports for the 1866 epidemic began coming in, Farr (1868) was very rapidly able to determine that residents in the catchment area of the East London Waterworks Company were disproportionately affected. He instigated an investigation of the company, which revealed that some of its water was being illegally pumped from the Old Ford reservoir on a polluted section of the River Lea. This exposure forced the company to discontinue using water from this source, whereupon the epidemic abated rapidly. This success more or less definitively confirmed the waterborne theory in the minds of most contemporary physicians, and it is accepted to this day.

## 4. Assessing the miasma theory

For the miasma theory in nineteenth-century England to count as a counterexample for deflationary realism, it must be demonstrated, firstly, that this theory enjoyed empirical success and, secondly, that the proposition that disease spreads *via* atmospheric miasmas was essential to this success. If these conditions are met, then the fact that this proposition is not preserved in successor theories violates the requirement of scientific continuity and thus refutes DR. In this section, it will be argued that, despite *prima facie* appearances, the miasma theory was *not* empirically successful and therefore fails as a counterexample to DR.

It may be argued the large-scale construction of sewage systems in the mid-nineteenth century, including Bazalgette's in London, is the result of a use-novel prediction of the miasma theory. It follows quite naturally from the basic principles of the theory that reducing humans' proximity to faecal matter would reduce the incidence of cholera, and yet this disease had certainly not previously been successfully controlled by these means. And the reformers responsible for these construction projects were, at least in their early stages, largely motivated by the miasma theory. Moreover, it seems like the theoretical prediction was successful – the dramatic decline in cholera and other diseases after the 1860s is largely attributable to infrastructure improvements of this type.

Furthermore, applying the unification criterion of empirical success more directly, it appears that the miasma theory genuinely unifies several lower-level empirical regularities. This unity is manifested, for instance, in Chadwick's report of 1842. Although the report is, for obvious reasons, rather focussed on the problem of cholera, it also mentions malaria in connection with poor drainage and the presence of rotting material:

> "… I visited the district, and examined the cottages and families living there. The land is nearly on a level with the water, the ground is marshy, and the sewers all open. Before reaching the district, I was assailed by a most

> disagreeable smell; and it was clear to the sense that the air was full of most injurious malaria" (Chadwick, 1842, s. 1)

Although the disease we now classify as malaria was indeed endemic to parts of England from the mid-17th until the beginning of the 20th century, primarily in marshy lands on the southeast and east coasts[21], it should be made clear that contemporary medical practice did not distinguish it from other fevers with broadly similar symptoms. Chadwick himself refers to "malaria" in connection with areas where it is extremely unlikely that it ever existed (Dobson, 1989; Hutchinson & Lindsay, 2006). Nevertheless, it is clear that these apparently very different diseases were thought of as operating by broadly the same mechanism.

It is undeniable that human beings do, indeed, tend to become ill in the presence of various kinds of rotting organic matter. Fevers like malaria and yellow fever are usually contracted around swamps and marshes. The presence of faecal matter in the environment leads to cholera and typhoid. And the bubonic plague, among many other diseases, visibly spreads more rapidly to homes that contain food waste and are generally dirty. What is challenging in all this for the realist is that this commonality is, from our modern perspective, completely "accidental". The mode of disease transmission in each case is different: malaria is transmitted by the bites of the mosquitoes that live in the swamp; cholera by the drinking of contaminated water; and bubonic plague by the bites of the fleas that live on rats – and these rats happen to be attracted to food leftovers.

The most convincing realist response to these challenges, following Vickers' (forthcoming) discussion of Velikovsky's theory (Chapter 2.10) and the achromatic telescope (Chapter 3.8) is to point out simply that the alleged "predictions" of the miasma theory are extremely vague. The various diseases that are allegedly explained by this theory are all extremely different in presentation, mortality, and

[21] It was eventually eradicated by draining of the marshes. Interestingly, however, there is no evidence that this was an intentional goal of these land-use changes – marshes were drained primarily in order to turn land over to more economically productive tilling or pasturage (Packard 2007, pp. 45-53).

pattern of spread. The theory gives no natural account of why the miasmas produced by faecal pollution should have such different effects to those produced by marshes, nor why faecal miasmas may give rise to cholera at some times and typhoid at others. This immediately suggests that the miasma theory is simply being used as a *post hoc* hypothesis to give some mechanistic explanation of an observed regularity.

And, in the case of cholera and sanitary reform, where the miasma theory *was* used to derive more specific predictions, these were disastrously incorrect. It had been observed, accurately, that ‘proximity’ to human faecal matter increased the chances of contracting this disease. The miasma theory predicted specifically that it was contact with the *airborne* emanations from this pollution that caused illness. Hence, to prevent the disease, the most important goal was removing this noxious material from people’s homes as soon as possible. This led, eventually, to the construction of advanced sewage systems which directed human waste far downstream of London. However, it initially led to the installation of flush toilets and the associated sewers which emptied directly into the parts of the Thames from which drinking water was drawn. It is doubly unfortunate that, facilitated by the water companies, the use of this water was increasing, at the expense of groundwater from wells, at just the same time that it was becoming more polluted. But as demonstrated by the Chadwick quote above, the use of piped water was *also* promoted by the miasma theorists! So, although these trends were no doubt also encouraged by the sheer convenience of piped water and flush toilets, the influence of the miasma theory probably increased the severity of the mid-century cholera epidemics. Thus, the miasma theory was hardly empirically successful in mid-nineteenth century Britain, and this case therefore does not pose a significant challenge to scientific realism.

## 5. The phlogistion theory and the chemical revolution

The phlogiston theory is included in many of the above authors’ lists. This is because, firstly, it was widely accepted by chemists during the eighteenth century and is reckoned to have enjoyed some degree of empirical success. Secondly, however, it is now believed that one of the central posits of this theory, namely that there exists a substance matching the properties ascribed to phlogiston, is false. This poses an obvious *prima facie* challenge to the deflationary realist, and hence to the partial realist and partial rationalist.

This challenge will be addressed from the perspective of the “chemical revolution” of the late eighteenth century, during which the phlogiston theory was largely superseded by Lavoisier’s oxygen theory. This section will give a broadly chronological outline of the various experimental results and theoretical innovations which, from the mid-eighteenth century onwards, combined to support the oxygen over the phlogiston theory in the minds of most chemists. As the goal of this section is to determine whether this transition is compatible with deflationary realism, particular attention will be paid to instances where either theory enjoyed empirical success. Since several authors writing about this episode have understood empirical success by reference to the UN account, this convention will generally be followed here. It will be taken as read, however, that a theory which is (not) successful under the UN is also (not) successful under the UV. Except where otherwise stated, the historical details in this section are drawn from Musgrave (1976), Hoyningen-Huene (2008) and Chang (2012). For ease of comprehension to the modern reader, the first time that contemporary chemical jargon is used to describe a substance or process, the equivalent modern term will also be given in square brackets.

The core ideas of phlogiston theory originated with the writings of Becher and Stahl in the late seventeenth and early eighteenth century. Phlogiston was understood to be an “imponderable fluid” or “principle” which can combine with ordinary matter. It affects the observable properties of the substances in which it is found, and is able to move from one substance to another. Moreover, certain substances, like charcoal [mostly carbon] and sulfur, contain large amounts of it, whereas other substances, such as metal ores, also called calxes [metal oxides], and ordinary air, contain relatively little. In general, phlogiston moves from substances that have a lot of it to those that have little, and this is what occurs in many chemical transformations. For instance, in combustion, a phlogiston-rich substance, such as sulphur, gives up phlogiston to the air. Another example is metal smelting, in which phlogiston is transferred from charcoal to an ore to produce the corresponding metal. One early predictive success of the phlogiston theory, described by Carrier (1993), was achieved by Stahl in 1697. It was known that the burning of sulphur and phosphorus in air resulted in vitriolic [sulphuric] and phosphoric acids, respectively). Since these acids were formed, Stahl reasoned, by the giving up of phlogiston to the air, the original materials should be recovered when these acids are heated in the presence of phlogiston-rich charcoal. This prediction, in fact, was correct.

In 1766, Cavendish produced “inflammable air” [hydrogen gas] by the reaction of metallic zinc, iron and tin with sulphuric and hydrochloric acids. Cavendish described inflammable air as extremely rich in phlogiston and in fact initially *identified* it with phlogiston. At least two novel predictions followed from this suggestion. The first was made by Cavendish as an immediate result of these experiments. He reasoned that, since the reaction of phlogiston-rich metal with acid results in a salt plus phlogiston-rich inflammable air, the reaction of the phlogiston-poor calx with an acid should yield the salt only. The second prediction was made by Priestly in 1783. Priestley reasoned that, since metal ores can be reduced to the corresponding metal by reaction with phlogiston-rich charcoal, they should also be so reduced by inflammable air, consuming the air in the process. Both of these predictions were borne out by experiment.

One major problem with phlogiston theory is that, if phlogiston is understood as a substance with mass, the release of phlogiston should (all else being equal) be associated with reduced weight. But many solid substances are observed to *increase* in weight as they undergo combustion or calcination [oxidation] in air. Phlogiston theorists offered several attempted remedies to this inconsistency in the middle part of the eighteenth century, including (infamously) the idea that phlogiston has negative weight. No solution was generally considered to be satisfactory, however, and the problem was largely left to one side. Lavoisier, however, took this problem sufficiently seriously that he proposed to overthrow the phlogiston theory altogether. In 1772, he proposed that combustion and calcination involve the *combining* of the solid substance with air, rather than the emission of phlogiston. To account for the fact that combustion always ceases when only a fraction of a trapped volume of air has been depleted, he argued that it was only "the purest part" of the air that was involved in this process. He later named this "oxygen".

Applying this new oxygen theory, Lavoisier reinterpreted the phenomena of metal calcination and smelting. Since calcination is the combination of the purest part of the air with the metal, it follows that the reversal of calcination should result in the release of this "pure air" along with the formation of the metal. But it was known that the reduction of metal calxes, as in smelting, typically resulted in "fixed air" [carbon dioxide], which has very different properties to ordinary air. Lavoisier explained this by proposing that fixed air was in fact a combination of pure air and carbon from the charcoal usually used in smelting. So, if a calx could be reduced to a pure metal in the absence of charcoal, it is predicted that the result would be pure oxygen. This matches exactly an experimental result obtained by Priestly in 1774. Priestley found that heating mercury calx resulted in metallic mercury and what he called "dephlogisticated air". This new form of air also supported combustion and respiration[22] better than normal air, matching exactly the properties that Lavoisier would expect to find for oxygen.

Lavoisier's prediction of this result was, incidentally, not "novel" by any intuitive standard (see Cook & Lauer, 1968; Emsley, 2001, pp. 297–304 for an account of the discovery of oxygen/dephlogisticated air). Scheele was the first to isolate the gas, in 1772, and Priestly is considered the discoverer only because he published his results first. Moreover, and more importantly, Lavoisier was almost certainly

[22] Priestly found that mice could survive for longer in an enclosed volume of this air than in an equal volume of ordinary air and, after sampling it himself, remarked that "[t]he feeling of it to my lungs was not sensibly different to that of common air; but I fancied that my breast felt peculiarly light and easy for some time afterwards" (Priestly, 1774,  s. 5)

aware of Priestley's discovery, since Priestly had visited him in France shortly after Priestley had performed the experiment but before Lavoisier made the prediction. However, the prediction still counts as an empirical success under either the use-novelty account or the unification accounts – it is sufficient that the result follows naturally from the stated theoretical principles, even if this consequence was not actually worked out until Lavoisier heard of Priestley's work.

Another significant episode came in 1783, when Cavendish discovered that, if inflammable air and dephlogisticated air are exploded together, the result is pure water. Oxygen and phlogiston theorists interpreted this result very differently. Lavoisier inferred that water is in fact a compound of the elements oxygen and inflammable air (which he renamed "hydrogen", meaning "water producing"). This led Lavoisier to several successful novel predictions. Firstly, he predicted that the reduction of a calx in hydrogen would also result in water, since oxygen was also available in the calx. Moreover, and perhaps more impressively, he inferred that, since the calcination (i.e. rusting) of iron or other metals in water involves combination with oxygen, it would produce pure hydrogen gas as a by-product. These predictions were duly confirmed experimentally.

At the same time, Cavendish's 1783 discovery led him to propose a new version of phlogiston theory. He proposed that dephlogisticated air was none other than dephlogisticated *water* and inflammable air phlogisticated water. This easily explains the initial observation, as one would expect the excess and deficit phlogiston to cancel out when the two combine, leaving only water. This new theory also aimed, at least in broad qualitative terms, to account for the changes in mass that are seen in various chemical reactions and that motivated Lavoisier's theory. And, as pointed by Musgrave (1976, p. 204), it was relatively successful in this respect. For instance, in calcination, the metal gives up its phlogiston to the dephlogisticated air, forming water, and this water combines with the metal to form a calx heavier than the metal. When inflammable air (now phlogisticated water) combines with a metal calx, the phlogiston from the air combines with the metal and drives out the water that was associated with it, reducing the weight of material. The water that is observed to form originates partly from the air and partly from the calx. And the formation of this water from the reduction of a metal calx is predicted independently by both Cavendish and Lavoisier.

So, by 1783, both Lavoisier's oxygen theory and Cavendish's new formulation of the phlogiston theory explained the same core set of phenomena. Nevertheless, chemists began to defect *en masse* to the oxygen theory from around this time, and it was almost entirely dominant by the beginning of the nineteenth century. The challenge is to determine whether this episode is compatible with the continuity of science that is asserted by deflationary (and partial) realism and by partial rationalism. This task is tackled in the following section

## 6. Assessing the chemical revolution

Musgrave claims that the victory of the oxygen theory is attributable to its empirical success under the UN account. Regarding the phlogiston theory, in contrast, Musgrave cites with approval Lavoisier's remark of 1783:

> "Chemists have made a vague principle of phlogiston which is not strictly defined, and which in consequence accommodates itself to every explanation into which it is pressed... It is a veritable Proteus which changes its form every minute." (quoted in Musgrave, 1976, p. 203)

The phlogiston theory, in other words, was no longer able to produce novel predictions. Instead, its proponents were simply making *ad hoc* modifications to accommodate the phenomena. It was a "degenerating research programme" (*ibid*, p. 203).

In assessing Musgrave's argument, it is worth distinguishing two propositions. Firstly, one might claim that the phlogiston theory was *not at all* empirically successful. Although broad in scope, it was only ever able to provide *post hoc* explanations of phenomena, and so never made any novel predictions. Secondly, one might accept that phlogiston theory did, in fact, achieve *some* use-novel predictions, but maintain that the oxygen theory is *more* empirically successful overall. Thus, applying Worrall's conception of predictive success as the basic theoretical virtue, it would be rational to accept the oxygen over the phlogiston theory.

The first proposition suggested above is clearly not tenable, in light of the historical evidence Musgrave himself presents. The phlogiston theory in fact produced several novel predictions, noted in the historical outline above. And, although it is true that the phlogiston theory was modified in response to unexpected empirical results, this was not done in a way characteristic of a degenerating research programme. After all, the oxygen theorists *also* modified their theory, and sometimes in response to the same empirical results! For instance, Cavendish's discovery of 1783 prompted both him and Lavoisier to alter their respective theories. In each case, alteration to the theory yielded verified empirical consequences going beyond the observation that prompted the change. Both actually predicted that water would be produced in the reduction of a calx by inflammable air. Moreover, while only Lavoisier actually made the prediction, it also follows from Cavendish's version of the phlogiston theory that inflammable air will result from the rusting of iron in water. So phlogiston theory was predictively successful up to and including the period where Lavoisier's oxygen theory was gaining dominance. This refutes Musgrave's claim that phlogiston was a degenerating research programme.

Moreover, from the perspective of modern chemistry, there is a very good explanation of the empirical success of phlogiston theory. Specifically, there is a strong structural similarity between the notions of phlogistication-dephlogistication on the one hand, and the modern concepts of reduction-oxidation on the other. This point has been made recently in the philosophical literature by Post (1971), Schurz (2008) and Ladyman (2009), among others. This similarity was also noticed some time ago by professional chemists:

> “If [the phlogistonists] had only thought to say ‘The substance burning gives up its phlogiston to, and then combines with, the oxygen of the air’, the phlogiston theory would never have fallen into disrepute. Indeed, it is curious now to note that not only their new classification but even their mechanism was essentially correct. It is only in the last few years that we have realized that every process that we call reduction or oxidation is the gain or loss of an almost imponderable substance, which we do not call phlogiston but electrons.” (Lewis, 1926, pp. 167–168)

As an example, consider the simple combustion reaction between charcoal and oxygen. A modern chemist would say of this reaction that the oxygen is reduced and the charcoal oxidised, or equivalently that electron density is transferred from charcoal to oxygen. A phlogiston theorist, analogously, would say that phlogiston is transferred from charcoal to dephlogisticated air.

All the substances that were classified as “phlogiston-rich” or “phlogiston-poor” are now thought of as electron-donors/reducing agents or electron-receivers/oxidising agents, respectively. The two classificatory schemes are structurally similar, and generate similar (and accurate) predictions about which substances will react with which. The most striking example of this is Priestly's successful prediction that a calx can be reduced to the corresponding metal by the action of inflammable air. He was able to make this prediction because he posited some basic chemical similarity between inflammable air and charcoal (which was already known to reduce metal calxes). And, when we understand phlogiston as playing a similar structural role to electron density in modern chemistry, we see that this posit was (from our perspective) essentially correct.

The second proposition suggested above, that the oxygen theory is *overall* more empirically successful than the phlogiston theory, is more plausible. However, it is worth pointing out that the oxygen theory, as described by Lavoisier and his contemporaries, did not preserve all the elements essential to the success of the phlogiston theory. The oxygen theory provides tools for tracking the movement of elements through chemical processes, but does not offer any systematic account of *why* substances react with each other. So it cannot be argued simply that the oxygen theory had all the predictive success of the phlogiston theory, *plus* some additional success. The situational is much more akin to the proverbial comparison of apples and oranges. Assuming that some sensible comparison can be made, however, the Worrall-style rationalist could reasonably argue that the choice of the phlogiston over the oxygen theory was rationally defensible.

The approach of treating empirical success as simply a theoretical virtue which theories may possess to a greater or lesser degree is, however, fundamentally at odds with the realist *motivation* for this form of rationalism advocated by Worrall, and with partial realism itself. The puzzle, from this perspective, is not merely that phlogiston theory is predictively successful, but that it achieves this success by positing a classificatory system which is completely absent from the oxygen theory. Moreover, this cannot be dismissed as a “happy accident” because, from the perspective of modern chemistry, this classificatory system appears to be roughly accurate.

Partial rationalism predicts that the dispute between phlogiston and oxygen theories would be resolved by a synthesis that incorporates the essential elements of each theory. Modern chemistry does represent such a synthesis. Theoretical concepts derived from Lavoisier’s oxygen chemistry identify the basic “building blocks” of substances, namely the chemical elements. Moreover, the idea that samples of these elements are of constant mass provides effective means for tracking their various rearrangements into different chemical compounds. Concepts that are structurally similar to those of phlogiston chemistry successfully categorise different chemical substances by their redox potential, accounting for the transfer of electrons and thereby (at least in many cases) explaining *why* a given reaction takes place.

However, it appears that the discipline of chemistry in the immediate aftermath of the chemical revolution does *not* represent a synthesis of these concepts as modern chemistry does. It is true that various electrochemical experiments were conducted around the period in question, and phlogiston theorists interpreted these in ways that modern chemists would regard as quite natural if phlogiston is identified with electrons. This history is surveyed by Allchin (1992) and Chang (2012, ch. 2). Nevertheless, it is undeniable that the phlogiston theory went into decline, and the abstract system of chemical classification associated with it passed out of use. This system only returned when thermodynamics had developed to the extent that it could be integrated with electrochemistry, during the latter part of the nineteenth-century. There was, in other words, an ‘interregnum’ during which concepts that contributed essentially to the empirical success of the phlogiston theory were not incorporated in the dominant theoretical framework in chemistry.

Partial rationalism is committed both to the existence of objectively rational standards for theory choice, and to the claim that most or all actual scientists involved in episodes of theory change conform to this standard (though not necessarily self-consciously). The existence of an ‘interregnum’ in the history of chemistry is thus potentially disastrous for this view. Assuming that this interpretation of historical events is accurate, there are three broad ways in which one might respond.

Firstly, one might take the failure of the descriptive part of the thesis as representing a major counterexample to rationalism in general, and thus opt for the Kuhnian (or some other non-rationalist) view of the history of science. Indeed, the proposition that certain essential elements fail to be preserved across theory-change fits well with the concept of “Kuhn loss”. Kuhn himself discusses the example of the chemical revolution in these terms:

> “[A]s disseminated in the nineteenth century, Lavoisier's chemical theory inhibited chemists from asking why the metals were so much alike, a question that phlogistic chemistry had both asked and answered. The transition to Lavoisier's paradigm had, like the transition to Newton's, meant

> a loss not only of a permissible question but of an achieved solution. That loss was not, however, permanent either. In the twentieth century questions about the qualities of chemical substances have entered science again, together with some answers to them." (Kuhn, 1962/1996, pp. 148-149)

In the light of the criteria of empirical success discussed in this thesis, the phlogistonist explanation of why the metals are "so much alike" is in fact not a particularly impressive achievement. That certain (not all) highly phlogisticated substances are shiny, malleable, etc. is certainly an interesting empirical regularity. But the proposition that this regularity exists, and that certain substances conform to it, is not entailed by any of the basic posits of the phlogiston theory. It thus does not contribute significantly to the unity or the degree of predictive success of the theory. The diagnosis of Kuhn loss is, however, more plausible in respect of the phlogiston system's predictions regarding chemical reactivity.

Chang (2012) has argued that the chemical revolution should be viewed as part of a larger trend in chemistry, namely the ascent of "compositionism" at the expense of "principlism". Principlism, of which phlogiston theory is a subtype, understands chemical reactions as the *transformation* of basic elements by the addition or removal of special substances called "principles". Mass is not necessarily conserved, but is simply another property which may be transformed by a reaction. The most important experimental evidence for principlist theories involves the *qualitative* features of substances. In contrast, compositionism understands chemical reactions as *rearrangements* of basic constituents. It follows that the most important sort of experimental evidence is found in precise *quantitative* measurements of mass, which reflects the distribution of these constituents. Chang argues that it is the growing dominance of these *experimental* practices associated with compositionism that drove the chemical revolution, rather than any theoretical argument in favour of Lavoisier's view.

Even if Chang's particular explanation for the decline of the phlogiston in favour of the oxygen theory is unsatisfactory, the anti-realist is undoubtedly in a good position to argue that the rationalist explanation is insufficient. And, if rationalism is unable to account for the chemical revolution, *some* ‘Kuhnian’ explanation involving non-rational factors must be invoked. However, as pointed out in Chapter 4.3, the Kuhnian view is incompatible with the cases where theoretical synthesis *does* occur. So it would seem that neither type of account is able to account for all historical episodes.

The second possible response to the descriptive failure of partial rationalism is to argue that this is simply an indication that the epistemic standards of this particular form of rationalism are incorrect. A variant of rationalism that applies a more adequate set of standards would, presumably, not encounter such counterexamples from the history of science. This type of argument is not particularly plausible, however, as epistemic standards for scientific reasoning, at least at this fundamental level, must surely be established on *a priori* grounds. In particular, the basic architecture of partial rationalism is fixed by the requirements of the NMA, and any normative standard that picks out suitable targets for this argument will face similar restraints.

The third possible response is to hold that the normative standards of partial rationalism *are* appealing on *a priori* grounds, and maintain that scientists collectively made a poor decision under the specific circumstances of the chemical revolution. The major prediction of partial rationalism in this case – that the essential concepts of both phlogiston and oxygen theories should be retained in a merged theory – was, after all, sustained in the end. So, even if DR, and thus partial rationalism, fails as a descriptive thesis, the normative/methodological component of these theses might nevertheless be vindicated by this example. Notice also, that this example does not challenge the claim of the partial realist that the essential elements of the phlogiston theory accurately represent the world – merely the secondary claim that such elements will be preserved in *immediate* successor theories.

This argument can be reinforced by recalling the point made in Chapter 1.9, that the suggested normative or methodological dimension is more strongly supported by cases where scientists ‘unintentionally’ achieved significant continuity between older and newer theories. This is significant in this case because the fact that there was a conceptual “interregnum” between the phlogiston theory of chemistry and modern electrochemistry seems to rule out the possibility that selection was involved in achieving the structural similarities between them.

Moreover, even if this counterexample does weaken DR as a normative prescription, it need not utterly discredit it. As pointed out in Chapter 1.7, the deflationary realist retains the “fallback option” of maintaining that the continuity of science is sustained in the vast *majority* of cases. This is even compatible with the NMA, as this does not, in fact, require that there are *no* miracles, only that they are exceedingly rare. The viability of this idea does depend, however, on both the *a priori* plausibility of partial rationalism and the actual prevalence of disconfirming cases in the history of science. This thesis has provided a fair number of arguments for the former, but has barely skimmed the surface of the latter problem. It may, therefore, be the case that there are further counterexamples to DR.

## **7. Chapter summary**

In this chapter, two representative cases have been selected from the lists of putative counterexamples to scientific realism suggested by Laudan, Stanford, Lyons and Vickers. To impose some analytical clarity on the discussion, these cases have been examined specifically as counterexamples to the claim which I have named deflationary realism. As DR is a logical consequence of both partial rationalism and partial realism, these more substantive views are also potentially threatened by these case studies.

Although the miasma theory does account for some observed regularities in the prevalence of disease, it does not do so with sufficient precision to be counted as empirically successful for the purposes of DR. This lack of empirical success is manifested in the fact that application of the miasma theory very likely worsened the cholera epidemics which struck London in the nineteenth century.

The phlogiston theory is a more difficult case, in that it *is* empirically successful by the standards of the UN account and the UV. Because the essential explanatory concepts of this theory were lost to chemistry, but were eventually recovered, this case stands as neither a confirming instance nor a simple counterexample to DR. Arguably, it vindicates DR as a methodological prescription, but not as a descriptive account of the history of science. The history of this case is compatible with the view that the NMA (properly applied) is a sound means for reasoning about scientific theories, and that scientists blunder when they fail to apply it. Even this view might be challenged, however, if further counterexamples emerge from the history of science.

## Thesis summary

The major arguments respectively for and against scientific realism are the no-miracles argument (NMA) and the pessimistic meta-induction (PMI). Following the trend in much of the scientific realism literature, I have argued that the best response to these arguments is to endorse some version of partial realism. This position states that, if a theory is empirically successful and some element of this theory is essential for that success, then we have good reason to believe (i) that *this element* accurately describes a corresponding feature of the world and (ii) that this element will be preserved in successor theories. A major aim of this thesis has been to articulate a defensible version of this view.

The aim of Chapter 2 was to articulate a defensible interpretation of the notion of “empirical success”. Currently the most common approach in the literature is to identify empirical success with one or other kind of novel predictive success. This approach represents what Musgrave calls a “historical” approach to theory confirmation. All these accounts accept, at least tacitly, that the extent to which a particular empirical result confirms (counts towards the empirical success of) a theory depends on the historical context in which it is introduced. In this chapter, it was argued that no historical account of confirmation can give an *absolute* account of empirical success. And an absolute account is required to warrant the inference under the NMA that the successful theory is (possibly only partly and/or approximately) *true*. Thus a *logical* account of theory confirmation is required, one which considers only the two-place logical relation between theory and a body of verified empirical results.

It has been argued that, of the accounts of empirical successful proposed in the literature, only those which emphasise the unifying power of a theory satisfy the demand for a two-place logical relation. These are grouped under the title of “weak predictivism” because, although they deny that novel prediction is a fundamental epistemic virtue of theories, these accounts do imply that novel prediction is an *indicator* that a theory is unifying. A particular version of weak predictivism is developed from Worrall’s use-novelty account of predictive success and named the “unification view” (UV). The final version of the UV articulated above states that a theory is confirmed just in case, and to the extent that, it precisely gives rise to more verified empirical content than that required to construct it, and does not give rise to excessive falsified content. This view is defended on the grounds that the major motivating intuition behind the NMA is the idea that there is something right about a theory where “you get out more than you put in”, and that this intuition is captured by the notion of unifying power.

Chapter 3 has addressed the question of what parts of an empirically successful theory are essential to that success. The NMA is the major motivation behind partial realism and this argument, if valid, licences the inference from empirical success to those propositions that, if true, explain that success. It has therefore been argued that whatever elements are picked out by the favoured account of essentialness must be sufficient to *explain* the empirical success of the theory. However, at least if explanation is understood in the model of deductive entailment, it is too easy to select a set of theoretical posits that count as explanatory. Indeed, since the mere statement of an empirical phenomenon entails itself, a simple-minded application of the suggested criterion could result in an account which regards only such statements as essential. A satisfactory account of essentialness must therefore pick out those elements which *fully* explain the empirical success of the relevant theory, but also exclude those elements that do not add any explanatory value.

Applying this criterion, several accounts, namely direct reference theories and structural realism, are rejected as being too ‘inclusive’. Entity realism and phenomenological realism, on the other hand, are rejected as too ‘exclusive’. The working posits account, and the closely related *divide et impera* strategy, are subject to many different interpretations. It is argued, however, that all fail to satisfy the stated criterion in various ways. The causal interpretation is, again, too inclusive. Vickers’ idea of selecting the “essentially contributing parts” of “derivation internal posits” is the most promising interpretation. However, because it focuses only on the derivations of individual results and on relations of deductive entailment between posits, it is also too exclusive.

The view of essentialness articulated and defended in this thesis is explicitly modelled on the account of empirical success advocated in Chapter 2 and is named the "empirically successful sub-theory account" (ESSA). The ESSA states that the essential posits of theories are those that give rise to more verified empirical results and/or empirically successful lower-level posits than required to construct them. This view satisfies the major criterion for an account of essentialness because the posits it selects, if (approximately) true together unify the accurate empirical claims of the theory and so explain its empirical success.

A second major aim of this thesis has been to defend deflationary realism (DR) as a minimal, "core" realist position. Because DR is a logical consequence of partial realism, any refutation of the former is also a refutation of the latter. So any assessment of DR is also an assessment of more metaphysically committed forms of realism. Moreover, even if a metaphysically committed form of realism is rejected on principled grounds, as it is by the constructive empiricist, DR can still be accepted as an empirical generalisation over the history of science. And the acceptance of this generalisation has normative consequences – if it has always been the case that there is continuity in science, scientists have very strong (inductive) grounds to seriously entertain only those new theories which preserve the essential elements of the older theories. This is especially so if there are many cases where this continuity has been achieved without any intervention by scientists specifically to select new theories that preserve the conceptual apparatus of older theories.

The third major aim of this thesis has been to assess DR against some putative counterexamples in the history of science. Along the way, in Chapter 4, the version of realism defended in this thesis was used as a means of illuminating the ongoing debate around the rationality of scientific "revolutions". If, as suggested in Chapter 2, unifying power ought to be the primary criterion for accepting a theory as true, this certainly has implications for cases of theory choice. The standard for believing that a theory is *true* is surely stronger than that for merely preferring it over rival theories! This slogan captures the basic flavour of the realist argument for rationalism in theory choice. A position based on partial realism as opposed to naive realism, however, states merely that, during episodes of theory change, scientists are rational to accept those *parts* of theories responsible for empirical successes (and have largely done so). A consequence of this partial rationalist position is that that it is sometimes rational to accept a *synthesis* of several competing theories, rather than choose between them. If this claim is sustained, it directly refutes Kuhn's view, as he strongly emphasises the "incommensurability" of competing scientific paradigms. The partial rationalist view was found to give adequate grounds for accepting the Copernican model of the universe over the Ptolemaic model, and thus successfully account for the occurrence of the Copernican "revolution". More importantly, however, several instances of theory change, namely the "neo-Darwinian" synthesis and the "prion revolution", were shown to be cases of theoretical synthesis.

Cases with a stronger *prima facie* claim to be counterexamples to DR were examined in Chapter 5. The two particular cases selected were the miasma theory of disease in nineteenth century England, and the phlogiston theory of chemistry in the eighteenth century. It was shown that, contrary to initial impressions, the miasma theory was not in fact empirically successful and so does not represent a challenge to DR. The phlogiston theory, in contrast, *was* empirically successful. And, although modern electrochemistry contains concepts that are structurally similar to the essential posits of the phlogiston theory, these posits were *not* preserved in the phlogiston theory's immediate successor. It consequence of this, it is argued, DR does not offer a universally accurate description of the history of science. The phlogiston theory is only a single counterexample, however, so it is difficult to assess how frequently DR is inaccurate. More importantly, however, the normative/methodological dimension of DR is strongly supported by this case. Chemists would, at least potentially, have attained some of the empirical success of modern electrochemistry at an earlier stage if they had preserved the essential elements of the phlogiston theory. Moreover, scientists did not deliberately craft modern electrochemistry so that it preserved the concepts of phlogiston theory (since the latter was not considered a salient reference point by the time the modern theory was constructed). This can be viewed as strong evidence for the general claim that scientists are well-advised, when devising new theories, to restrict themselves to theories that preserve the essential elements of successful earlier theories.

## References

1. Ackerknecht, E. H. (1948/2009). Anticontagionism between 1821 and 1867: The Fielding H. Garrison Lecture. *International journal of epidemiology*, *38*(1), 7–21.
2. Ainsworth, P. M. (2009). Newman's Objection. *The British Journal for the Philosophy of Science*, *60*, 135–171.
3. Airy, G. (1931/1842). *Mathematical tracts on the lunar and planetary theories, 3rd edition*. Cambridge: Macmillan and Co.
4. Allchin, D. (1992). Phlogiston after oxygen. *Ambix*, *39*(3), 110–116.
5. Barnes, B., & Bloor, D. (1982). Relativism, rationalism and the sociology of knowledge. In M. Hollis & S. Lukes (Eds.), *Rationality and relativism (*pp. 21–47). Oxford: Basil Blackwell.
6. Barnes, B., D., & Bloor, J. H. (1996). *Scientific Knowledge: A Sociological Analysis*. Chicago: University of Chicago Press.
7. Bloor, D. (1976). *Knowledge and Social Imagery*. London: Routledge & Kegan Paul.
8. Bolton, D., & Bendheim, P. (1988). A modified host protein model of scrapie. In G. Bock & J. Marsh (Eds.), *Novel Infectious Agents and the Central Nervous System, Ciba Foundation Symposium 13* (pp. 164–181). Chichester, UK: John Wiley and Sons.
9. Bowler, P. J. (2009). *Evolution : the history of an idea*. Berkeley: University of California Press.
10. Boyd, R. (1980). Scientific realism and naturalistic epistemology. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 613–662.
11. Boyd, R. (1983). On the current status of the issue of scientific realism. *Erkenntnis*, *19*, 45–90.
12. Boyd, R. (1985). Lex orandi est lex credendi. In P. M. Churchland & C. A. Hooker (Eds.), *Images of science: Essays on realism and empiricism*. Chicago: University of Chicago Press.
13. Boyd, R. (1989). What realism implies and what it does not. *Dialectica*, *43*(1), 5–29.
14. Braithwaite, R. B. (1968). *Scientific Explanation a Study of the Function of Theory, Probability and Law in Science*. Cambridge: Cambridge University Press.
15. Brooker, G. (2008). Diffraction at a single ideally conducting slit. *Journal of Modern Optics*, *55*(3), 423–445.
16. Brown, J. (1982). The miracle of science. *The Philosophical Quarterly*, *32*(128), 232–244.
17. Butterfield, H. (1931/1965). *The Whig Interpretation of History*. New York: W. W. Norton & Co.
18. Carnap, R. (1928/1967). *The Logical Structure of the World*. London: Routledge & Kegan Paul.
19. Carp, R., Merz, P., Kascsak, R., Mertz, G., & Wisniewski, H. (1985). Nature of the scrapie agent: Current status of facts and hypotheses. *Journal of general virology*, *66*, 1357–1368.

20. Carrier, M. (1993). What is right with the miracle argument: establishing a taxonomy of natural kinds. *Studies in the History and Philosophy of Science*, *24*(3), 391–409.
21. Carroll, L. (1895). What the Tortoise Said to Achilles. *Mind*, *4*(14), 278–280.
22. Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Oxford University Press.
23. Cartwright, N. (1999a). Models and the limits of theory: quantum Hamiltonians and the BCS model of superconductivity. In M. Morgan & M. Morrison (Eds.), *Models as mediators: perspectives on natural and social science* (pp. 241–281). Cambridge: Cambridge University Press.
24. Cartwright, N. (1999b). *The Dappled World: A Study in the Boundaries of Science*. Cambridge University Press: Cambridge.
25. Cartwright, N. (2009). Entity Realism versus Phenomenological Realism versus High Theory Realism. London School of Economics: Scientific Realism Revisited Conference.
26. Cartwright, N., Shomar, T., & Suarez, M. (1995). The Tool Box of Science: Tools for the Building of Models with a Superconductivity Example. In W. E. Herfel (Ed.), *Theories and models in scientific processes: Proceedings of AFOS '94 Workshop, August 15-26, Madralin and IUHPS '94 Conference, August 27-29, Warszawa*. Amsterdam: Rodopi.
27. Cassirer, E. (1910/1953). *Substance and Function, and Einstein's Theory of Relativity.* New York: Dover Publications.
28. Cavendish, H. (1798). Experiments to Determine the Density of the Earth. *Philosophical Transactions of the Royal Society of London*, *88*, 469–526.
29. Cei, A. (2009). Structural Realism as a Form of Humility. In M. Suárez, M. Dorato, & M. Redei (Eds.), *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association* (pp. 35–45). Dordrecht: Springer.
30. Chadwick, E. (1842). Report on the sanitary condition of the labouring population and on the means of its improvement. Retrieved from http://www.deltaomega.org/ChadwickClassic.pdf (30th Sept. 2012)
31. Chakravartty, A. (1998). Semirealism. *Studies in the History and Philosophy of Science*, *29*(3), 391–408.
32. Chakravartty, A. (2003). The Structuralist Conception of Objects. *Philosophy of Science*, *70*(5), 867–878.
33. Chakravartty, A. (2004). Structuralism as a form of scientific realism. *International Studies in the Philosophy of Science*, *18*(2-3), 151–171.
34. Chakravartty, A. (2007). *A metaphysics for scientific realism: Knowing the unobservable*. Cambridge: Cambridge University Press.
35. Chakravartty, A. (2011). Scientific Realism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from http://plato.stanford.edu/archives/sum2011/entries/scientific-realism/ (30th Sept. 2012)
36. Chang, H. (2003). Preservative Realism and Its Discontents: Revisiting Caloric. *Philosophy of Science*, *70*, 902–912.
37. Chang, H. (2012). *Is Water H2O?: Evidence, Pluralism and Realism,*. Dordrecht: Springer.

38. Clarke, S. (2001). Defensible Territory for Entity Realism. *The British Journal for the Philosophy of Science*, *52*(4), 701–722.
39. Clegg, B. (2008). *Light years: an exploration of mankind's enduring fascination with light*. London: Macmillan.
40. Cook, G. A., & Lauer, C. M. (1968). Oxygen. In C. A. Hampel (Ed.), *The Encyclopedia of the Chemical Elements* (pp. 499–512). New York: Reinhold Book Corporation.
41. Copernicus, N. (1543/1978). De Revolutionibus (On the Revolutions). In E. Rosen (Ed.). Baltimore: Johns Hopkins University Press.
42. Cordero, A. (2011). Scientific Realism and the Divide et Impera Strategy: The Ether Saga Revisited. *Philosophy of Science*, *78*(5), 1120–1130.
43. Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 138–163.
44. Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*, 561–563.
45. Cruse, P. (2005). Ramsey sentences, structural realism and trivial realization. *Studies In History and Philosophy of Science Part A*, *36*(3), 557–576.
46. Cruse, P., & Papineau, D. (2002). Scientific realism without reference. In M. Marsonet (Ed.), *The problem of realism* (pp. 174–189). Aldershot, UK: Ashgate.
47. Cummiskey, D. (1992). Reference failure and scientific realism: a response to the meta-induction. *The British journal for the philosophy of*, *43*(1), 21–40.
48. Delisle, J.-N. (1715). Reflexions. *Mémoires de l'Académie Royale des Sciences* (pp. 166–169). Paris: S. Landry & au Griffon.
49. Demopoulos, W., & Friedman, M. (1985). Bertrand Russell's the analysis of matter: Its historical context and contemporary interest. *Philosophy of Science*, *52*(4), 621–639.
50. Devitt, M. (1984). *Realism and Truth*. Oxford: Blackwell.
51. Dickinson, H. (1954). *The Water supply of Greater London*. London: Newcomen Society.
52. Dobson, M. J. (1989). History of malaria in England. *Journal of the Royal Society of Medicine*, *82*(17), 3–7.
53. Duhem, P. (1906/1954). *The Aim and Structure of Physical Theory*. Princeton, New Jersey: Princeton University Press.
54. Dyson, F. W., Eddington, a. S., & Davidson, C. (1920). A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *220*, 291–333.
55. Eddington, A. (1939). *The Philosophy of Physical Science*. Cambridge: Cambridge University Press.
56. Elsamahi, M. (2004). A critique of localized realism. *Philosophy of Science*, *72*(5), 1350–1360.
57. Emsley, J. (2001). *Nature's Building Blocks: An A-Z Guide to the Elements*. *Nature's Building Blocks: An A-Z Guide to the Elements*. Oxford: Oxford University Press.
58. Eyler, J. M. (1973). William Farr on the cholera: the sanitarian's disease theory and the statistician's method. *Journal of the history of medicine and allied sciences*, *28*(2), 79–100.

59. Eyler, J. M. (2001). The changing assessments of John Snow's and William Farr's cholera studies. *Sozial- und Präventivmedizin SPM*, *46*(4), 225–232.

60. Farr, W. (1852). *Report on the mortality of cholera in England, 1848-49.* London: W. Clowes.

61. Farr, W. (1868). *Report on the cholera epidemic of 1866 in England, Supplement to the 29th Annual Report of the Registrar-Geneneral*. London: George E. Eyre and William Spottiswoode.

62. Feyerabend, P. (1975/1993). *Against Method*. London: Verso.

63. Fine, A. (1984). The natural ontological attitude. In Jarrett Leplin (Ed.), *Scientific Realism* (pp. 83–107). Berkeley: University of California Press.

64. Fine, A. (1986a). Unnatural Attitudes: Realist and Instrumentalist Attachments to Science. *Mind*, *95*(378), 149–179.

65. Fine, A. (1986b). *The Shaky Game: Einstein, Realism, and the Quantum Theory*. Chicago: University of Chicago Press.

66. Fine, A. (1991). Piecemeal realism. *Philosophical Studies*, *61*(1), 79–96.

67. Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.

68. Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

69. Forster, M. R. (1988a). Sober's Principle of Common Cause and the Problem of Comparing Incomplete Hypotheses. *Philosophy of Science*, *55*(4), 538–559.

70. Forster, M. R. (1988b). Unification, explanation, and the composition of causes in Newtonian mechanics. *Studies In History and Philosophy of Science Part A*, *19*(1), 55–101.

71. French, S. (2003). Scribbling on the blank sheet: Eddington's structuralist conception of objects. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, *34*(2), 227–259.

72. French, S. (2009). Keeping quiet on the ontology of models. *Synthese*, *172*(2), 231–249.

73. Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, *71*(1), 5–19.

74. 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.

75. Gajdusek, C. (1985). Subacute Spongiform Virus Encephalopathies Caused by Unconventional Viruses. In K. Maramorosch & J. J. McKelvey (Eds.), *Subviral Pathogens of Plants and Animals: Viroids and Prions* (pp. 483–544). New York: Academic Press.

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

77. Goldman, A. I. (1979). What is Justified Belief? In G. S. Pappas (Ed.), *Justification and Knowledge: New Studies in Epistemology* (pp. 1–23). Dordrecht: Reidel.

78. Goldman, A. I. (1986). *Epistemology and Cognition*. Cambridge, Massachusetts: Harvard University Press.

79. Goodman, N. (1983). *Fact, Fiction and Forecast, 2nd Edition*. Indianapolis: Harvard University Press.
80. Gould, S. (2002). *The structure of evolutionary theory*. Cambridge, Massachusetts: Belknap Press.
81. Gower, B. (2000). Cassirer, Schlick and “Structural” Realism: the Philosophy of the Exact Sciences in the Background To Early Logical Empiricism. *British Journal for the History of Philosophy*, *8*(1), 71–106.
82. Griffith, J. (1967). Self-replication and scrapie. *Nature*, *215*(5105), 1043–1044.
83. Hacking, I. (1982). Experimentation and scientific realism. *Philosophical Topics*, *13*(1), 71.
84. Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.
85. Hacking, I. (1988). Philosophers of experiment. *PSA: Proceedings of the Biennial Meeting of the* Philosophy of Science Association, 147–156. Hacking, I. (1989). Extragalactic reality: The case of gravitational lensing. *Philosophy of Science*, *54*(3), 327–350.
86. Halliday, S. (1999). *The great stink of London: Sir Joseph Bazalgette and the cleansing of the Victorian Capital*. Stroud: Sutton Publishing.
87. Hardin, C. L., & Rosenberg, A. (1982). In defense of convergent realism. *Philosophy of Science*, *49*(4), 604–615.
88. Hardy, A. (1984). Water and the search for public health in London in the eighteenth and nineteenth centuries. *Medical history*, *28*(3), 250–282.
89. Hardy, A. (1991). Parish pump to private pipes: London’s water supply in the nineteenth century. *Medical History*, *Supplement No. 11*, 76–93.
90. Harker, D. (2008). On the Predilections for Predictions. *The British Journal for the Philosophy of Science*, *59*(3), 429–453.
91. Harman, G. (1965). The inference to the best explanation. *The Philosophical Review*, *74*(1), 88–95.
92. Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.
93. Hippocrates. (1978). Airs, Waters, Places. In G. E. R. Lloyd (Ed.), *Hippocratic Writings* (pp. 148–169). London: Penguin.
94. Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1–34.
95. Howson, C. (1984). Bayesianism and support by novel facts. *The British Journal for the Philosophy of Science*, *35*(3), 245–251.
96. Howson, C. (1990). Fitting your theory to the facts: Probably not such a bad thing after all. In C. W. Savage (Ed.), *Minnesota studies in the philosophy of science* (pp. 224–244).
97. Howson, C. (2000). *Hume’s Problem: Induction and the Justification of Belief*. Oxford: Oxford University Press.
98. Hoyningen-Huene, P. (2008). Thomas Kuhn and the chemical revolution. *Foundations of Chemistry*, *10*(2), 101–115.
99. Hutchinson, R. A., & Lindsay, S. W. (2006). Malaria and deaths in the English marshes. *Lancet*, *367*, 1947-1951

100. Jardine, N. (2003). Whigs and stories: Herbert Butterfield and the historiography of science. *History of Science*, *41*, 125–140.
101. Jenkin, F. (1867). Review of The origin of species. *North British Review*, *46*, 277–318.
102. Ketland, J. (2004). Empirical Adequacy and Ramsification. *The British Journal for the Philosophy of Science*, *55*(2), 287–300.
103. Keyes, M. (1999a). The Prion Challenge to the “Central Dogma” of Molecular Biology, 1965-1991: Part I: Prelude to Prions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *30*(1), 1–19.
104. Keyes, M. (1999b). The Prion Challenge to the “Central Dogma” of Molecular Biology, 1965-1991: Part II: The Problem with Prions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *30*(2), 181–218.
105. Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.
106. Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. C. Salmon (Eds.), *Scientific Explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.
107. Kitcher, P. (1993). *The Advancement of Science*. Oxford: Oxford University Press.
108. Kripke, S. A. (1980). *Naming and necessity*. Oxford: Basil Blackwell.
109. Kuhn, T. S. (1957). *The Copernican revolution: Planetary astronomy in the development of Western thought*. Cambridge Massachusetts: Harvard University Press.
110. Kuhn, T. S. (1977). Objectivity, Value Judgement, and Theory Choice. *The essential tension* (pp. 320–339). Chicago: University of Chicago Press.
111. Kuhn, T. S. (1962/1996). *The Structure of Scientific Revolutions, 3rd Edition*. Chicago: University of Chicago Press.
112. Kyburg, H. (1961). *Probability and the Logic of Rational Belief*. Middletown, CT: Wesleyan University Press.
113. Ladyman, J. (1999). Review. A novel defense of scientific realism. Jarrett Leplin. *The British Journal for the Philosophy of*, *50*, 181–188.
114. Ladyman, J. (2002). *Understanding Philosophy of Science*. London: Routledge.
115. Ladyman, J. (2009). Structural realism versus standard scientific realism: the case of phlogiston and dephlogisticated air. *Synthese*, *180*(2), 87–101.
116. Ladyman, J., & Lipton, P. (2006). Wouldn’t it be lovely: Explanation and Scientific Realism. *Metascience*, *14*(3), 331–361.
117. Lakatos, I. (1968). Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, *69*, 149–186.
118. Lange, M. (2001). The apparent superiority of prediction to accommodation as a side effect: a reply to Maher. *The British journal for the philosophy of science*, *52*, 575–588.
119. Laudan, L. (1981). A Confutation of Convergent Realism. *Philosophy of Science*, *48*(1), 19–49.
120. Laudan, L. (1984). Realism without the real. *Philosophy of Science*, *51*(1), 156–162.

121. Leplin, J. (1997). *A novel defense of scientific realism*. Oxford: Oxford University Press.
122. Lewens, T. (2006). Science Undermined by Our Limited Imagination? *Science*, *313*(August), 1047–1048.
123. Lewis, G. N. (1926). *The anatomy of science*. New Haven, Connecticut: Yale University Press.
124. Lewis, P. J. (2001). Why the pessimistic induction is a fallacy. *Synthese*, *129*(3), 371–380.
125. Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, 247–266.
126. Lipton, P. (1993). Is the best good enough? *Proceedings of the Aristotelian Society*, *93*(1993), 89–104.
127. Lipton, P. (1994). Truth, existence, and the best explanation. In A. A. Derksen (Ed.), *The Scientific Realism of Rom Harré*, (pp. 89–110). Tilburg: Tilburg University Press.
128. Lipton, P. (2004). *Inference to the Best Explanation, 2nd Edition*. London: Routledge.
129. Longmate, N. (1966). *King Cholera: the biography of a disease*. London: Hamish Hamilton.
130. Lyons, T. D. (2002). Scientific Realism and the Pessimistic Meta-Modus Tollens. In Steven Clarke & T. Lyons (Eds.), *Recent Themes in the Philosophy of Science: Scientific Realism and Commonsense* (pp. 63–90). Dordrecht: Kluwer.
131. 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.
132. Magnus, P. D., & Callender, C. (2004). Realist Ennui and the Base Rate Fallacy. *Philosophy of Science*, *71*(3), 320–338.
133. Maher, P. (1988). Prediction, accommodation, and the logic of discovery. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 273–285.
134. Maher, P. (1990). How Prediction Enhances confirmation. In J. M. Dunn, A. Gupta, & N. D. Belnap (Eds.), *Truth Or Consequences: Essays in Honor of Nuel Belnap* (pp. 327–344). Dordrecht: Kluwer.
135. Maher, P. (1993). Howson and Franklin on prediction. *Philosophy of Science*, *60*(2), 329–340.
136. Mahon, B. (2003). The Man Who Changed Everything: The Life of James Clerk Maxwell. Chichester, UK: John Wiley & Sons
137. Maraldi, G. F. (1723). Diverses expèriences d'optique. *Mémoires de l'Académie Royale des Sciences* (pp. 111–143). Paris: S. Landry & au Griffon.
138. Maxwell, G. (1962). Theories, Frameworks, and Ontology. *Philosophy of Science*, *29*(2), 132–138.
139. Maxwell, G. (1970a). Structural Realism and the Meaning of Theoretical Terms. In M. Radner & S. Winokur (Eds.), *Analysis of Theories and Methods of Physics and Psychology* (pp. 181–192). Minneapolis: University of Minnesota Press.

140. Maxwell, G. (1970b). Theories, Perception, and Structural Realism. In R. G. Colodny (Ed.), *The Nature and Function of Scientific Theories* (pp. 3–34). Pittsburgh: University of Pittsburgh Press.
141. Maxwell, J. C. (1878). Ether. In T. S. Baynes (Ed.), *Encyclopædia Britannica Ninth Edition, vol. 8* (pp. 568 – 572). Edinburgh: A. & C. Black.
142. Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, *58*(4), 523–552.
143. Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
144. Mayo, D. G. (2008). How to Discount Double-Counting When It Counts: Some Clarifications. *The British Journal for the Philosophy of Science*, *59*(4), 857–879.
145. Mayo, D. G. (2010). 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* (pp. 155–169). Cambridge: Cambridge University Press.
146. McCormmach, R. (1998). Mr. Cavendish weighs the world. *Proceedings of the American Philosophical Society*, *142*(3), 355–366.
147. McKinley, M., Bolton, D., & Prusiner, S. (1983). A protease-resistant protein is a structural component of the scrapie prion. *Cell*, *35*(1), 57–62.
148. McLeod, K. S. (2000). Our sense of Snow: the myth of John Snow in medical geography. *Social science & medicine (1982)*, *50*(7-8), 923–35.
149. Morrison, M. (1998). Modelling Nature: Between Physics and the Physical World. *Philosophia Naturalis*, *38*, 65–85.
150. Morrison, M. (1999). Models as Autonomous Agents. In M. Morgan & M. Morrison (Eds.), *Models as Mediators. Perspectives on Natural and Social Science.* (pp. 38–65). Cambridge: Cambridge University Press.
151. Musgrave, A. (1974). Logical versus Historical Theories of Confirmation. *The British Journal for the Philosophy of Science*, *25*(1), 1–23.
152. Musgrave, A. (1976). Why did oxygen supplant phlogiston?: Research programmes in the chemical revolution. In C. Howson (Ed.), *Method and appraisal in the physical sciences: the critical background to modern science, 1800-1905* (pp. 181–210). Cambridge University Press.
153. Musgrave, A. (1988). The ultimate argument for scientific realism. In R. Nola (Ed.), *Relativism and realism in science* (pp. 229–252). Dordrecht: Kluwer.
154. Nelson, D. L., & Cox, M. M. (2005). *Lehninger Principles of Biochemistry, 4th Edition*. New York: W.H. Freeman.
155. Newman, M. (2005). Ramsey-Sentence Realism as an Answer to the Pessimistic Meta-Induction. *Philosophy of Science*, *72*, 1373–1384.
156. Newman, M. (2010). Beyond Structural Realism: pluralist criteria for theory evaluation. *Synthese*, *174*(3), 413–443.
157. Newman, M. H. A. (1928). Mr. Russell's "Causal Theory of Perception." *Mind*, *37*(146), 137–148.
158. Nickles, T. (1987). Lakatosian heuristics and epistemic support. *The British journal for the philosophy of science*, *38*(2), 181–205.
159. Niiniluoto, I. (1999). *Critical Scientific Realism*. Oxford: Oxford University Press.

160. Nozick, R. (1974). *Anarchy, State, and Utopia*. Oxford: Blackwell.
161. Oesch, B., Groth, D., Prusiner, S., & Weissmann, C. (1988). Search for a Scrapie-Specific Nucleic Acid: A Progress Report. In G. Bock & J. Marsh (Eds.), *Novel Infectious Agents and the Central Nervous System, Ciba Foundation Symposium 13* (pp. 209–223). Chichester, UK: John Wiley and Sons.
162. Oesch, B., Westaway, D., Walchli, M., McKinley, M., Kent, S., Aebersold, R., Barry, R., *et al.* (1985). A Cellular Gene Encodes Scrapie PrP 27–30 Protein. *Cell*, *40*, 735–746.
163. Packard, R. M. (2007). *The making of a tropical disease: a short history of malaria*. Baltimore: Johns Hopkins University Press.
164. Papineau, D. (1992). Reliabilism, induction and scepticism. *The Philosophical Quarterly*, *42*(166), 1–20.
165. Papineau, D. (1993). *Philosophical Naturalism*. Oxford: Basil Blackwell.
166. Papineau, D. (2010). Realism, Ramsey sentences and the pessimistic meta-induction. *Studies In History and Philosophy of Science Part A*, *41*(4), 375–385.
167. Pattison, I. (1982). Scrapie a "gene"? *Nature*, *299*(5880), 200.
168. Peirce, C. S. (1958). *The collected works of Charles Sanders Peirce*. Cambridge, Massachusetts: Harvard University Press.
169. Poincaré, H. (1902/1952). *Science and Hypothesis*. New York: Dover Publications, Inc.
170. Popper, K. (1963/2002a). *Conjectures and Refutations: the Growth of Scientific Knowledge*. London: Routledge Classics.
171. Popper, K. (1952/2002b). *The logic of scientific discovery*. London: Routledge Classics.
172. Porter, R. (1999). *The greatest benefit to mankind: a medical history of humanity*. London: Fontana Press.
173. Post, H. (1971). Correspondence, invariance and heuristics: in praise of conservative induction. *Studies In History and Philosophy of Science Part A*, *93*(3), 213–255.
174. Priestley, J. (1774). *Experiments and observations on different kinds of air, Volume 2*. London: J. Johnson.
175. Prusiner, S. (1982). Novel proteinaceous infectious particles cause scrapie. *Science*, *216*(4542), 136–144.
176. Prusiner, S. (1984). Prions. *Scientific American*, *251*, 50–59.
177. Prusiner, S., Bolton, D., Groth, D., Bowman, K., Cochran, S., & McKinley, M. (1982). Further purification and characterization of scrapie prions. *Biochemistry*, *21*(26), 6942–6950.
178. Prusiner, S., McKinley, M., Groth, D., Bowman, K., Mock, N., Cochran, S., & Masiarz, F. (1981). Scrapie agent contains a hydrophobic protein. *Proceedings of the …*, *78*(11), 6675–6679.
179. Prusiner, S., Scott, M., Foster, D., Pan, K., Groth, D., Mirenda, D., Torchia, M., *et al.* (1990). Transgenetic studies implicate interactions between homologous PrP isoforms in scrapie prion replication. *Cell*, *63*, 673–686.
180. Psillos, S. (1994). A Philosophical Study of the Transition from the Caloric Theory of Heat to Thermodynamics : Resisting the Pessimistic Meta-

Induction. *Studies In History and Philosophy of Science Part A*, *25*(2), 159–190.

181. Psillos, S. (1995). Is Structural Realism The Best of Both Worlds? *Dialectica*, *49*(1), 15–46.
182. Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.
183. Psillos, S. (2001). Is structural realism possible? *Philosophy of Science*, *68*(3), 13–24.
184. Psillos, S. (2006). Thinking about the ultimate argument for realism. In C. Cheyne & J. Worrall (Eds.), *Rationality and Reality: Conversations with Alan Musgrave* (pp. 133–156). Dordrecht: Springer.
185. Putnam, H. (1973). Meaning and reference. *The Journal of Philosophy*, *70*(19), 699–711.
186. Putnam, H. (1975a). *Philosophical papers, Vol. 1, Mathematics, matter and method.* Cambridge: Cambridge University Press.
187. Putnam, H. (1975b). *Philosophical papers, Vol. 2, Mind, Language and Reality*. Cambridge: Cambridge University Press.
188. Putnam, H. (1978). *Meaning and the Moral Sciences*. London: Routledge and Kegan Paul.
189. Quine, W. V. O. (1961). Two Dogmas of Empiricism. *From a Logical Point of View* (pp. 20–46). Cambridge, Massachusetts: Harvard University Press.
190. Ramsey, F. P. (1931). Theories. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics, and Other Logical Essays* (pp. 212–236). London: Routledge & Kegan Paul.
191. Reichenbach, H. (1949). *The Theory of Probability*. Berkeley: University of California Press.
192. Reiner, R., & Pierson, R. (1995). Hacking's experimental realism: an untenable middle ground. *Philosophy of Science*, *62*(1), 60–69.
193. Reiter, P. (2000). From Shakespeare to Defoe: malaria in England in the Little Ice Age. *Emerging infectious diseases*, *6*(1), 1–11.
194. Resnik, D. (1994). Hacking's experimental realism. *Canadian Journal of Philosophy*, *24*(3), 395–412.
195. Romer, M., & Cohen, I. B. (1940). Roemer and the First Determination of the Velocity of Light (1676). *Isis*, *31*(2), 327–379.
196. Ronchi, V. (1957). *Optics : the science of vision*. New York: New York University Press.
197. Roush, S. (2009). Optimism about the Pessimistic Induction. In P. D. Magnus & J. Busch (Eds.), *New Waves in Philosophy of Science* (pp. 29–58). London: Palgrave Macmillan.
198. Russell, B. (1919). *Introduction to Mathematical Philosophy*. London: George Allen & Unwin.
199. Russell, B. (1927/1992). *The Analysis of Matter*. London: Routledge.
200. Ryle, G. (1949). *The concept of mind*. London: Hutchinson.
201. Saatsi, J. (2009). Scientific Realism and Historical Evidence: Shortcomings of the Current State of Debate. *EPSA Philosophy Of Science: Amsterdam 2009*, 329–340.
202. Saatsi, J., Psillos, S., Winther, R. G., & Stanford, P. K. (2009). Review Symposium: Grasping At Realist Straws. *Metascience*, *18*(3), 355–390.

203. Saatsi, J., & Vickers, P. (2010). Miraculous Success? Inconsistency and Untruth in Kirchhoff's Diffraction Theory. *The British Journal for the Philosophy of Science*, *62*(1), 29–46.
204. Scerri, E., & Worrall, J. (2001). Prediction and the periodic table. *Studies In History and Philosophy of Science Part A*, *32*(3), 407–452.
205. Schlick, M. (1918/1974). *General Theory of Knowledge*. Vienna: Springer-Verlag.
206. Schurz, G. (2008). When Empirical Success Implies Theoretical Reference: A Structural Correspondence Theorem. *The British Journal for the Philosophy of Science*, *60*(1), 101–133.
207. Shapin, S., & Schaffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle and the Experimental Life*. Princeton, New Jersey: Princeton University Press.
208. Shomar, T. (1998). *Phenomenological Realism, Superconductivity and Quantum Mechanics.* Thesis at the University of London.
209. Shomar, T. (2008). Phenomenologism vs fundamentalism: The case of superconductivity. *Current Science*, *94*(10), 1256–1264.
210. Shomar, T. (2009). Bohr as a Phenomenological Realist. *Journal for General Philosophy of Science*, *39*(2), 321–349.
211. Smart, J. C. (1963). *Philosophy and scientific realism*. London: Routledge and Kegan Paul.
212. Snow, J. (1849). *On the mode of communication of cholera*. London: John Churchill.
213. Snow, J. (1855). *On the mode of communication of cholera*. London: John Churchill.
214. Sober, E. (2001). What is the problem of simplicity? In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference and modeling: Keeping it sophisticatedly simple* (pp. 13–31). Cambridge: Cambridge University Press.
215. Stanford, P. K. (2003a). Pyrrhic victories for scientific realism. *The journal of philosophy*, *100*(11), 553–572.
216. Stanford, P. K. (2003b). No refuge for realism: Selective confirmation and the history of science. *Philosophy of Science*, *70*(5), 913–925.
217. Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives.* Oxford: Oxford University Press.
218. Suarez, M. (1999). The role of models in the application of scientific theories: Epistemological implications. In M. Morgan & M. Morrison (Eds.), *Models as Mediators. Perspectives on Natural and Social Science.* (pp. 168–196). Cambridge: Cambridge University Press.
219. Swenson, L. (1972). *The ethereal aether: a history of the Michelson-Morley-Miller aether-drift experiments, 1880-1930*. Austin: University of Texas Press.
220. Van Fraassen, B. (1989). *Laws and symmetry*. Oxford University Press Oxford.
221. Velikovsky, I. (1950). *Worlds in collision*. Garden City, New York: Doubleday & Co.
222. Vickers, P. (2011). Theory Eliminativism as a Methodological Tool, 1–20. Retrieved from http://philsci-archive.pitt.edu/8472/ (30th Sept. 2012)

223. Vickers, P. (2012). Historical magic in old quantum theory? *European Journal for Philosophy of Science*, *2*(1), 1–19.
224. Vickers, P. (forthcoming). The Nascency of the Divide et Impera Debate.
225. Whewell, W. (1840/1847). *Philosophy of the Inductive Sciences*. London: John W. Parker.
226. Whewell, W. (1858/1968). *William Whewell's Theory of Scientific Method*. (R. Butts, Ed.). Pittsburgh: University of Pittsburgh Press.
227. Worrall, J. (1984). An unreal image. *The British Journal for the Philosophy of Science*, *35*(1), 65–80.
228. Worrall, J. (1985). Scientific discovery and theory-confirmation. In J. C. Pitt (Ed.), *Change and progress in modern science*. Dordrecht: D. Reidel Publishing Company.
229. Worrall, J. (1989a). Structural Realism: The Best of Both Worlds? *Dialectica*, *43*, 99–124.
230. Worrall, J. (1989b). Fresnel, Poisson and the "White Spot": The Role of Successful Prediction in Theory-acceptance. In D. Gooding, T. J. Pinch, & S. Schaffner (Eds.), *The Uses of Experiment - Studies of Experimentation in Natural Science* (pp. 135–158). Cambridge: Cambridge University Press.
231. Worrall, J. (1990). Scientific revolutions and scientific rationality: The case of the "elderly holdout." In C. W. Savage (Ed.), *Scientific theories* (pp. 319–354). Minneapolis: University of Minnesota Press.
232. Worrall, J. (1994). How to Remain (Reasonably) Optimistic: Scientific Realism and the "Luminiferous Ether." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1994*, 334–342.
233. Worrall, J. (2000). The Scope, Limits, and Distinctiveness of the Method of "Deduction from the Phenomena": Some Lessons from Newton's "Demonstrations" in Optics. *The British Journal for the Philosophy of Science*, *51*(1), 45–80.
234. Worrall, J. (2002). New evidence for old. In P. Gardenfors, J. Wolenski, & K. Kijania-Placek (Eds.), *In the Scope of Logic, Methodology and Philosophy of Science* (pp. 191–209). Dordrecht: Kluwer Academic Publishers.
235. Worrall, J. (2005). Miracles, Pessimism and Scientific Realism. *LSE webpage*, (October 2005), 1–55. Retrieved from http://www2.lse.ac.uk/philosophy/WhosWho/staffhomepages/Publications/NMAandPIfinal.pdf (30th Sept. 2012)
236. Worrall, J. (2006). Theory-confirmation and history. In Colin Cheyne & J. Worrall (Eds.), *Rationality and Reality: Conversations with Alan Musgrave* (pp. 31–61). Dordrecht: Springer.
237. Worrall, J. (2007). Miracles and Models: Why reports of the death of Structural Realism may be exaggerated. *Royal Institute of Philosophy Supplements*, *82*(61), 125–154.
238. Worrall, J. (2010a). For universal rules, against induction. *Philosophy of Science*, *77*(5), 740–753.
239. Worrall, J. (2010b). Theory Confirmation and Novel Evidence: 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* (pp. 125–154). Cambridge: Cambridge University Press.

240. Worrall, J. (2011). Miracles and Structural Realism. In E. M. Landry & D. P. Rickles (Eds.), *Structural Realism: Structure, Object, and Causality* (pp. 77–98). Dordrecht: Springer.
241. Worrall, J., & Zahar, E. (2001). Appendix IV: Ramseyfication and Structural Realism. In E. Zahar (Ed.), *Poincare’s Philosophy: From Conventionalism to Phenomenology* (pp. 236–251). Chicago: Open Court.
242. Wray, K. B. (2007). A selectionist explanation for the success and failures of science. *Erkenntnis*, *67*(1), 81–89.
243. Wray, K. B. (2010). Selection and Predictive Success. *Erkenntnis*, *72*(3), 365–377.
244. Zahar, E. (1973). Why did Einstein’s Programme supersede Lorentz's? (I). *The British Journal for the Philosophy of Science*, *24*(2), 95–123.
245. Zahar, E. (2004). Ramseyfication and structural realism. *Theoria*, *49*, 5–30.
246. van Fraassen, B. (1980). *The Scientific Image*. Oxford, England: Oxford University Press.
247. von Helmholtz, H. (1899). Preface. In H. Hertz (Ed.), *The Principles of Mechanics Presented in a New Form*. London: Macmillan.