Addressing Kuhn's challenge: conceptual continuity and natural kinds

Birkbeck, University of London

Addressing Kuhn's Challenge:

Conceptual Continuity and Natural Kinds

Thesis for the degree of PhD

Magnus Fried

Abstract.

Thomas Kuhn poses a fundamental worry about explaining scientific progress, which I call *Kuhn's Challenge*. The Challenge consists of two related questions:

(A) If the meanings of key terms change between theories on either side of a paradigm shift, how can we still say that these theories are about the same thing?

(B) Even if we assume that two theories address the same subject matter, how can we determine which one is better?

A popular reply to Kuhn is to adopt a semantics for natural kind terms influenced by Kripke in *Naming and Necessity* and Putnam in "The Meaning of 'Meaning'", according to which such terms rigidly refer – independently of theory changes – to the same kinds across possible worlds and through time. I argue that this approach can explain extra-theoretical conceptual continuity only if we assume that all natural kinds have the same *essence type*. Though Kripke and Putnam take for granted that this essence type is microstructural, I argue that in practice, many sciences postulate natural kinds with other essence types, such as historical or functional essences; and that when new discoveries are made, prompting paradigm shifts, the relevant essence type may change. Moreover, which type is relevant to which science is as much a matter of decision as of discovery. Such a claim may seem to threaten realism about natural kinds. I argue, however, that we can be both pluralists and realists, if we recognise that conceptual continuity is secured *ex post*. Contrary to those who have argued for similar positions, I claim that we need not give up the rigidity of natural kind terms or the global ambitions of realism. In the end I show how the framework I have developed illuminates the debate over Kripke's argument against Physicalism in the philosophy of mind.

1 Contents

Introduction

The starting point for this thesis is the challenge for scientific progress, and by implication for scientific realism, that Thomas Kuhn formulates in *The Structure of Scientific Revolutions*. Intuitively, we would only want to describe a theory T' as representing progress from a theory T if both address the same subject matter, and if there is a method of measurement where T' scores better than T. But this is what Kuhn's analysis of the history of science challenges. In Chapter 1, I lay out Kuhn's Challenge, which I suggest is constituted by the following two questions:

> *(A) If the meanings of key terms change between theories on either side of a paradigm shift, how can we still say that these theories are about the same thing? And,*
>
> *(B) Even if we assume that two theories do address the same subject matter, how can we determine which one is the better?*

Though Kuhn's most radical claims cannot be sustained, I argue that Kuhn's Challenge remains a serious threat.

In Chapters 2 and 3, I introduce the Kripke-Putnam semantics, often taken to provide a response to Kuhn. I present Saul Kripke's semantic machinery for proper names, but also the extension to natural kind terms that Hilary Putnam develops in more detail. I argue that there is one major, relevant difference between them: Contrary to the usual interpretation, Kripke does *not* provide a semantic account that can be used to answer Kuhn, because his account presupposes that we are operating with our current language. Putnam, on the other hand, claims that his account addresses Kuhn, because it supports extra-theoretical conceptual continuity.

Putnam presents two types of argument, one based on historical chains and one on necessity. Both rely on two major assumptions: the validity of thought experiments and extra-theoretical essentialism. I conclude that the reliance on thought experiments is justified, but that the essentialism is problematic. Natural kinds need to have the same essence over time and across possible worlds for the arguments to work, which also means that there must have been continuity in terms of what I call their "essence type", the category to which the essence belongs. For Putnam in "The Meaning of 'Meaning'" this is not an issue, as he regards the microstructural essence type as the obvious choice for natural kinds.

Chapter 4 suggests, to the contrary, that the choice of essence type is far from obvious and discusses what reasons there could be to deviate from what I see as the default position for philosophy of science: acceptance of actual scientific practices. If we accept current practices, we see a variety of essence types – especially functional and historical essences, in addition to

microstructural essences – used to good effect to form theories with high explanatory power. This power exists only given a particular context, a purpose of enquiry.

Chapter 5 presents a series of arguments in favour of scientific realism and finds them wanting, partly because they rely on the defunct idea, which I call the "Perfect Theory Theory". This idea says that there in principle exists a theory the posits of which has a perfect and unique match to features in nature. This chapter also addresses the heterogeneity of actual sciences that might look like a threat to arguments for continuity and progress. But I argue that pluralism and realism are compatible, once the notion of a perfect theory is given up; indeed, realism without that notion should embrace pluralism. However, I identify a dependency of scientific realism on conceptual continuity, which still has to be resolved.

Chapter 6 starts with another apparent issue for conceptual continuity, namely the crucial role of decision-making. Many developments in the history of the sciences appear to be based on decisions rather than discoveries. I claim, however, that recognising decision-making provides the key to how Putnam's historical-chain argument for continuity can be defended. Conceptual continuity of natural kind terms is based on decisions taken for good reasons and can be described in well-justified *ex post* stories.

My framework is developed piece-by-piece, chapter-by-chapter, to aid my analysis. In Chapter 7, I pull the pieces together and apply them to the semantics of natural kind terms. In so doing, I provide a way to shore up Putnam's second argument for conceptual continuity, the necessity-based argument. Contrary to some theorists who also adopt pluralism, I defend the extension of the Kripke-Putnam machinery to natural kind terms. My approach, I propose, is compatible with their machinery – with one exception. The combination of the two points towards a more promising semantics for natural kind terms. I show how it allows good but seemingly conflicting arguments to be accommodated.

I continue the application of my framework in Chapter 8, which serves as a more elaborate proof-of-concept, by demonstrating how it can help to throw some light on the Mind-Body problem. In this chapter, the semantic tools developed previously are put within the context of Kuhn's phases of scientific development. I argue that the debate over the Mind-Body problem is best construed as having reached a scientific crisis. The result of this analysis is that we can recognise the force of Kripke's critique of Physicalism without giving up all hope for future scientific explanations of the human mind, by one or many theories.

1 Kuhn's Challenge

1.1 Introduction

Thomas Kuhn's main book, *The Structure of Scientific Revolutions* (*Structure*),[1] tells a story about how scientists, particularly physicists and chemists, learn and work within their respective scientific communities, and how these practices explain the success of their enterprise.

Kuhn was trained as a physicist, but became a science historian rather than a practising scientist. When he wrote *Structure*, Kuhn had already published a book on the Copernican Revolution. As a historian, Kuhn wants his model to fairly represent not just current but also past scientists, an ambition that has far-reaching consequences for his view of how scientific practices develop and change. In *Structure*, Kuhn takes a step further, into the philosophy of science, touching also on metaphysics and epistemology. He suggests that a correct historical analysis of scientific practices over time, understanding the reason why they have been so successful, has radical consequences for philosophy.

Kuhn's work on the proper understanding of earlier scientists, and the explanation as to why these scientists are so poorly understood in modern textbooks, leads him to conclusions about the nature of scientific development, and the relationship between old and new theories, that pose a serious challenge to a realist view of scientific progress. Some writers have elaborated this challenge, while others have tried to find counter-arguments.

One picture of scientific progress is this: science accumulates knowledge as each generation of scientists builds on the results of previous ones, towards an increasingly better understanding of nature. Old theories fail crucial tests and new theories are introduced to explain all that the old theories did, adding explanations for new cases, including the ones where those old theories failed. The Positivists describe scientific activity as an attempt to verify theories. Karl Popper turns this around and sees it as consisting of formulations of risky hypotheses (which he calls

[1] Kuhn [2012]. (All page references without further qualification are to this edition).

"conjectures") and attempts to falsify them (attempted "refutations").[2] For Popper, this is what (real) scientists actually do, and also what they *should* do to be good scientists.

In contrast to these earlier accounts, Kuhn presents a model that separates a normal science phase, in which scientists are guided by a *paradigm,* from an extra-ordinary (or *revolutionary*) phase, bridged by an interim period of scientific crisis; I will detail these below. In *Structure* Kuhn says that his theory can be seen as combining the two earlier schools. But in fact, it is different from both. During normal science, there is no attempt to either verify or falsify an established theory, while during revolutionary science, there is no accumulation of knowledge. The break in continuity in this revolutionary phase is the crucial point for the discussion about progress in science, to which I will keep returning.

In this thesis, I separate two related aspects of *Structure*. The first aspect, which I will call "Kuhn's sociological theses",[3] describes the actual behaviour of scientists, or rather, of scientific communities. Kuhn famously uses the term 'paradigm' to describe what guides scientists during stable, normal-science periods, but is subject to change with new discoveries and inventions during scientific revolutions.

The second aspect concerns the conclusions Kuhn draws for classical philosophical issues. There are exciting hints about radical implications for ontology, truth, and the philosophy of science. I will focus on those conclusions that are related to philosophical models for scientific progress, and refer to them as "Kuhn's Challenge". A central notion is *incommensurability* between paradigms, which threatens to stop any attempt to justify scientific progress dead in its tracks by implying that the differences and barriers between paradigms are such that we cannot be justified to speak of continuity, and therefore of progress, at all.

In contrast to his sociological work, Kuhn's views on the philosophical implications of this work are difficult to pin down. He oscillates between stronger and weaker theses already in *Structure*,

[2] See Popper [1959] and Popper [1989].

[3] In apparent disagreement with this, Ian Hacking In his Introductory Essay to the fourth edition of *Structure*, writes: "Notice that there is *no* sociology in the book" (xxxvi). But this might be terminology only, as the next sentence starts: "Scientific communities and their practices are, however, at its core".

and repeatedly revisits the same themes, particularly incommensurability, in later writings. In the postscript to *Structure*, written in 1969, seven years after the main text ("Postscript"), some of his views are noticeably more mellow. But the uncompromising views in *Structure* are both historically and intellectually interesting. I will only occasionally refer to Kuhn texts other than *Structure* and "Postscript".

In this chapter, §§1.2-1.5 outline the nature of Kuhn's Challenge, and §§1.6-1.9 discuss how to interpret Kuhn's own response. I argue that although many of Kuhn's conclusions are less radical than they appear, Kuhn's Challenge remains a serious threat to a philosophical justification of scientific progress.

1.2 Normal Science

During the normal science stage, activities in a scientific community are governed by a *paradigm*. Kuhn uses this word in many different ways,[4] but the two main uses in *Structure* are (i) a particular breakthrough result, an *exemplar*, which becomes a model for further activities, and (ii) a connected framework of "beliefs, values, techniques, and so on"[5] governing a science. Kuhn's considered view, expressed in "Postscript", is that the *exemplar* sense (i) is the deeper sense of 'paradigm', and "the central element of what I now take to be the most novel and least understood aspect of this book."[6] The primacy of exemplars over rules and theories is expressed also in the main text of *Structure*. The paradigm, Kuhn says, does not follow from rules in a theory, it is the other way around: the rules are abstracted from the paradigm. Scientists learn to follow the exemplar in their training to solve increasingly more difficult tasks in the same spirit, and thereby get to understand associated theories, rules and concepts. But for my purposes, it will not matter which sense is more basic, and I will not distinguish between sense (i) and sense (ii).

[4] Margaret Masterman [1970] counted 21 different uses.
[5] P.174.
[6] P.86. Kuhn in "Postscript" calls a paradigm in sense (ii) a "disciplinary matrix".

The paradigms Kuhn has in mind include the broad exemplars and theories by Aristotle, Ptolemy, Newton, Lavoisier and Einstein. But Kuhn also recognises exemplars that have a similar guiding role, but in a narrower field, such as Maxwell's equations.[7]

It is not the role of the scientists to question or try to falsify the paradigm, not as long as the period of normal science lasts, Kuhn claims. He also, contrary to Popper, holds that this practice is sound, and that it makes it possible for scientists to make fast progress within their discipline. It allows them to dedicate their time and efforts to solving specific issues, detailing and extending the use of the paradigm, without spending energy and time on questioning fundamentals. The paradigm helps to create a productive scientific community, insulated from the rest of society by its focus on more and more obscure and detailed studies. Kuhn compares normal science activities to puzzle-solving; it is always assumed that a solution exists, but it takes ingenuity to find it.

During normal science, the paradigm is taken for granted. Counter-instances always occur, but the scientific community regards them as tasks to work on – until the crisis arrives.

1.3 Crises in Scientific Communities

In any science, there are always many issues yet unexplained; these are the tasks for the puzzle-solvers. Indeed, it is one of the characteristics of a paradigm to be open-ended enough to provide a fruitful field for further research. But at one point, and for a variety of reasons, a set of unresolved issues becomes a bother, and pessimism sets in regarding the ability of the existing paradigm to solve them. This is the intermediary period, the scientific crisis. During this period, Kuhn states, there is no paradigm in either of the senses above, but several tentative, revolutionary ideas competing for success. Eventually, one of them is established as the new paradigm and the old one abandoned. Afterwards, when textbooks are rewritten from the perspective of the new paradigm, and a new set of puzzles to solve is established, the paradigm shift becomes almost invisible.

[7] Plausibly, theory changes affecting the meaning of concepts can occur more frequently and less dramatically outside physics and chemistry. I will ignore this difference when I discuss examples from different sciences.

As always, Kuhn is looking from a practising scientist's perspective. For a scientist, a crisis is an event where the scientific society loses faith in the paradigm. In the normal-science period, Kuhn says, the scientific community attributes failures to the scientist, and they do not reflect on the paradigm. "It is a poor carpenter who blames his tools", Kuhn quotes.[8] This situation changes with the crisis, when the community realises it needs retooling. It needs a new paradigm.

1.4 Finding a New Paradigm

Kuhn complicates the picture of when and by whom a new scientific discovery is made (we seldom know the actual discoverer, and never the exact time) and the relation between discoveries and inventions (there is no major difference). But revolutionary events, where new paradigms are introduced, nevertheless have the character of singularities compared to the steady progress of normal science. Kuhn stresses this difference. Scientists during normal science are introduced to theories and terms by practising on examples that follow the paradigm exemplar, but a paradigm shift introduces new theories and changes to key terms.

The explanation of the movement of planets highlights the point. Nicolaus Copernicus published his *De revolutionibus orbium coelestium* in 1543. It has since been regarded as a revolutionary contribution to science, putting the sun rather than the earth in the centre of the universe. But to compare Copernicus' theory with the preceding Ptolemaic theory is not straightforward, Kuhn shows. To say that Ptolemy believed that the planets move around the earth and that Copernicus proved that they move around the sun is, he argues, to say something confused. 'Planet' changed meaning with the new theory; for Ptolemy, the sun is one of the planets and cannot very well move around itself. 'The planets move around the sun' is not a statement that Ptolemy held to be false and Copernicus held to be true. According to Kuhn, Ptolemy would have regarded it as *meaningless*; it could neither be true nor false. To understand Kuhn's argument, I will mention one natural response and why it does not work.

The objection goes like this:

[8] P.80.

> We cannot count Ptolemaic astronomy as an alternative description, because it does not form a consistent whole with its data. As Kuhn himself remarks, the crisis was already there when Copernicus came along: the old theory had problems in fitting the data, and struggled both with explanation and prediction.[9] In the same way, the special creationist theory was already struggling in Darwin's days, as its assumptions did not form a consistent whole with relevant data.[10] An inconsistent theory, surely, can never be an alternative explanation.

This counter-argument misses the point. The bankruptcy of the old theories was indeed a fact when Copernicus and Darwin arrived on the scene, but this meant that the respective scientific communities had already left the normal-science mode and entered into crisis mode, giving up hope that the old theory would ever overcome its anomalies. In the period of normal science, this is not the case, as counter-instances, which occur for all theories, are treated as puzzles to solve, not anomalies. With the benefit of hindsight, one knows that some of these efforts will fail. But this knowledge was not available to the practitioners of normal science at the time; they were as justified as current scientists to believe in their paradigms and other tools. One scientist's puzzles are another (and later) scientist's fatal anomalies. Even less did they know which of their assumptions would eventually be given up, and what would replace them; more than one possibility existed when the choice was to be made. I will come back to this issue in §§1.7-1.8 where I discuss incommensurability.

How is a new paradigm chosen? It is not that a new paradigm explains everything that the old one explained and more; there are often phenomena explained by the old paradigm that are *not* covered by the new, Kuhn says. This can be the case because, for example, (a) some phenomena are excluded from the scope of this science due to increased specialisation, (b) the phenomena are no longer explained by the theory but included in or excluded by its axioms, or (c) the phenomena previously explained by the old paradigm now feature on the list for puzzle-solving.

[9] P.67: "The state of Ptolemaic astronomy was a scandal before Copernicus' announcement."
[10] This is not Kuhn's example. Darwin's theory is not mentioned in *Structure*, except as a metaphor (or possibly more) for non-teleological scientific development. But see LaPorte [2004] for an exploration of the species concept in Kuhn's spirit.

There are certainly always large overlaps, but "new paradigms seldom or never possess all the capabilities of their predecessors".[11] About Copernicus, Kuhn says that he "destroyed a time-honoured explanation of terrestrial motion without replacing it",[12] while Newton did the same for gravity.[13,]

It is not easy to produce any proof favouring one paradigm over the other, Kuhn says. Paradigm debates "are not really about relative problem-solving ability, though for good reasons they are usually couched in these terms."[14] And a "decision must be based less on past achievement than on future promise."[15] Persuasion sometimes works, but very often individual scientists never switch paradigms during their lifetimes.[16] Eventually, the balance in the scientific community tips in favour of the new paradigm when younger scientists join, and that paradigm becomes established, governing another era of normal science.

For each of Kuhn's examples, there are good reasons that led from the crisis to the acceptance of a new paradigm by the scientific community,[17] but not always the ones we might expect. Kuhn says that Copernicus' theory was neither simpler nor more accurate than its predecessor, but it offered hope for future research where the old, discredited theory offered none. This was vindicated when the increased precision provided by Johannes Kepler was added,[18] converting many astronomers to the new theory.

In some cases, the selling point can be that "the new paradigm permits the prediction of phenomena that had been entirely unsuspected while the old one prevailed."[19] Support for

[11] P.168. This is sometimes called a "Kuhn loss".
[12] P.156. Ptolemy had relied on Aristotle's Physics, but the heliocentric system lacked such underpinning before Galileo.
[13] P.105. Newton treated it as an "innate" attraction between particles, without a mechanical explanation, which Kuhn regards as reversing a scholastic standard, reversed again by Einstein.
[14] P.156.
[15] P.156.
[16] P.149: "How, then, are scientists brought to make this transposition? Part of the answer is that they are very often not."
[17] As opposed to individual scientists: Kepler was apparently drawn to Copernicus's heliocentric system for religious reasons, identifying the sun with the Father.
[18] In his laws of planetary motion.
[19] P.153.

Copernicus' approach materialised some 60 years after his death when improved instruments displayed new celestial details, such as mountains on the moon and the phases of Venus. As such details were consistent with Copernicus' approach, which treated these bodies as Earth-like, many non-astronomers now joined the converted.

The choice of a new paradigm is the result of an acceptance in the scientific community, Kuhn says. There is always a good reason for the choice, with the benefit of hindsight, but for Kuhn, the outcome is the result of a process for which there are no exceptionless criteria. "A decision of that kind can only be made on faith", he says, and "there is no single argument that can or should persuade them all".[20]

These claims have implications for how we should understand the relation between the earlier and the later paradigm. They create questions that must be addressed by an account for scientific progress, due to the conceptual changes that new paradigms introduce.

1.5 Conceptual Change and Its Problems

Revolutionary science produces new paradigms, which typically change the meaning of key terms. According to Kuhn, 'planet' does not mean the same for Ptolemy and Copernicus,[21] while 'mass'[22] and 'space'[23] do not mean the same for Aristotle, Newton or Einstein.

Scientists trained on the current paradigm do not necessarily have a good understanding of earlier scientists who followed another paradigm, because this is irrelevant for their job. But to understand scientific development over time, using a *historian's* perspective, Kuhn wants to understand these earlier scientists on their own terms. To do that, conceptual change has to be

[20] P.157.
[21] P.128: "[T]he Copernicans who denied its traditional title 'planet' to the sun were not only learning what 'planet' meant or what the sun was. Instead, they were changing the meaning of 'planet' so that it could continue to make useful distinctions in a world where all celestial bodies, not just the sun, were seen differently from the way they had been seen before."
[22] P.102: "Newtonian mass is conserved; Einsteinian is convertible with energy. Only at low relative velocities may the two be measured in the same way, and even then they must not be conceived to be the same."
[23] P.148: "What had previously been meant by space was necessarily flat, homogeneous, isotropic, and unaffected by the presence of matter. If it had not been, Newtonian physics would not have worked. To make the transition to Einstein's universe, the whole conceptual web whose strands are space, time, matter, force, and so on, had to be shifted and laid down again on nature whole."

taken into account. This leads to a fundamental worry about scientific development, or rather to two related questions, which together I label *Kuhn's Challenge*:

> (A) *If the meanings of key terms change between theories on either side of a paradigm shift, how can we say that these theories are about the same thing?* And,
> (B) *Even if we assume that two theories do address the same subject matter, how can we determine which one is better?*

Scientific progress within a paradigm is, for Kuhn, the steady resolution of puzzles, carrying out normal science. But to be able to describe progress in science across paradigm shifts, we need answers to both questions. My focus throughout this thesis is how to answer Kuhn's Challenge.

When discussing questions (A) and (B) it is essential to clarify that they relate to *philosophical models* of scientific progress, not to the fact of progress itself. I do not discuss whether scientific progress has actually taken place; that is the *explanandum*, what I in this thesis take as a given. This precisification is in line with Kuhn. Kuhn endorsed scientific progress to the extent that he saw an element of tautology involved: sciences are called "sciences" partly because they have been able to make such a remarkable progress, compared to other human activities.[24] My problem is to understand how this is epistemologically justified, given Kuhn's Challenge. I will start with Kuhn's own answer.

1.6 Kuhn's Response

How can Kuhn's Challenge be addressed? Kuhn's own analysis, where observations about the behaviour of scientists lead to conclusions in philosophy, seems to provide reasons to be pessimistic. But it is not always easy to determine exactly which conclusions Kuhn draws from his sociological material, since he sometimes leans towards using this material as illuminating metaphors, but at other times takes a stronger position.[25] I will try to separate claims about

[24] There is a complication: Kuhn, on p.168, also says the progress is judged from the perspective of the conquering, new paradigm, and that therefore: "The perception [of progress] is, in important respects, self-fulfilling."
[25] The misunderstandings Kuhn complains about in "Postscript" are therefore largely self-inflicted.

scientific communities from claims about scientific theories and concepts, and metaphors from literal statement.

One possible interpretation of *Structure* is that Kuhn answers the two questions in the negative, saying that there is no way to compare scientific theories, and therefore no way to claim that one is better than another, due to fundamental differences between their terms. That would make him a relativist, if we by that name mean somebody who holds that the value of a scientific result can only be judged relative to the reigning paradigm. According to this position, we cannot have an answer to either point (A) or point (B) and it makes no sense to talk about progress across paradigms.

To portray Kuhn as a relativist in *Structure* is not to construct a straw-man.[26] There are several claims that can be extracted from Kuhn's discussion that support a relativist interpretation:

> (a) There is no difference between discovery and invention. Kuhn says that he regards the distinction as artificial, and adds that this artificiality is "an important clue to several of this essay's main theses."[27]
>
> (b) It makes no sense to speak of scientific progress at all. "We must learn to recognize as causes what have ordinarily been taken to be effects. If we can do that, the phrases 'scientific progress' and even 'scientific objectivity' may come to seem in part redundant."[28]
>
> (c) Talk about progress lacks informative content. "Does a field make progress because it is a science, or is it a science because it makes progress?"[29]

[26] Although it is true that there often is something tentative in is his more radical statements about philosophical implications (as opposed to the sociological observations), indicated by the multitude of qualifiers: "we may want to say", "may come to seem in part", "must fail to make complete contact", "in important respects", "a sense in which", "not altogether inappropriate", etc.
[27] P.53.
[28] P.161.
[29] P.161.

(d) There are no objective standards to compare theories. "[T]here is no standard higher than the assent of the relevant community."[30]

(e) Truth is not a useful notion in the context of scientific development. "We may...have to relinquish the notion, explicit or implicit, that changes of paradigm carry scientists and those who learn from them closer and closer to the truth."[31]

(f) Talk about a world that is independent of our paradigm has no meaning. "[W]e may want to say that after a revolution scientists are responding to a different world."[32]

(g) The illusion of progress is due to the victor writing history. "Inevitably those remarks will suggest that a member of a mature scientific community is, like the typical character of Orwell's *1984*, the victim of a history rewritten by the powers that be. Furthermore, that suggestion is not altogether inappropriate."[33]

Central to Kuhn's metaphorical or literal relativism is his notion of *incommensurability*. The term sums up "several reasons why the proponents of competing paradigms must fail to make complete contact with each other's viewpoints".[34] Historians of science, Kuhn says in the beginning of *Structure*, discover "bodies of belief quite incompatible with the ones we hold today."[35] They therefore must realise that the history of science is not just one of steady accumulation. Historians should consequently not see earlier periods as contributing to later periods, but "attempt to display the historical integrity of that science in its own time."[36] Later he adds that new paradigms are "not only incompatible but often actually incommensurable with that which has gone before."[37] The differences are "both necessary and irreconcilable."[38]

[30] P.94.
[31] P.169.
[32] P.111.
[33] P.166.
[34] P.147.
[35] P.3.
[36] P.3.
[37] P.103.
[38] P.103.

These are strong words. Incommensurability is a serious threat to any discussion about scientific progress, leaving the philosopher the hopeless task of translating the untranslatable and comparing the incomparable. It conjures up a picture of science where monolithic theories are built and then abandoned by scientists, for reasons of their own, in favour of the next. It is also wide open to objections, for example that inter-disciplinary efforts continue to have considerable success. If such bridge-building is possible across disciplines, why not also across paradigms? And is not Kuhn himself an example of this in his books on the history of science, which seem perfectly possible to understand? I will consider translation issues first, and then what Kuhn regards as the most fundamental aspect of incommensurability: that scientists across paradigms "practise their trades in different worlds."[39]

1.7 Translation Issues

One aspect of incommensurability is the problem of translation across paradigms, caused by the changes in meaning of concepts. This sense of the term is primarily discussed in "Postscript", which in many ways conveys a much more nuanced picture than the main text.

In "Postscript", Kuhn notes that a communication breakdown between scientists of different paradigms cannot simply be overcome by stipulated definitions of troublesome terms, because this is not how the terms have been learned. Scientists learn terms in part from working on a series of tasks within a paradigm, making it difficult to extract the criteria for correct application. "They cannot, that is, resort to a neutral language which both use in the same way".[40] A lack of translation mechanism between paradigms relates to question (B) of Kuhn's Challenge: if we cannot translate, we seem to lack a basis for comparison. But it also relates to question (A): we would in that case have no way of *verifying* that the two paradigms refer to the same subject matter.

But there are, in Kuhn's considered opinion, at least partial remedies, which might mitigate this risk. The skills required are seldom available to the scientist, but if he does want to acquire them,

[39] P.149.
[40] P.200.

he needs to become a translator, where conceptual changes are taken into account to provide a fair description across paradigm gulfs. "That is what the historian of science regularly does (or should) when dealing with out-of-date scientific theories",[41] he writes in "Postscript". In his later "Dubbing and Redubbing", Kuhn compares translation between different scientific theories with the translation of fiction into another language:

> [T]he problems of translating a scientific text, whether into a foreign tongue or into a later version of the language in which it was written, are far more like those of translating literature than has generally been supposed. In both cases the translator repeatedly encounters sentences that can be rendered in several alternative ways, none of which captures them completely.[42]

Kuhn points to problems arising from conceptual changes to communication across gulfs, and to communication breakdowns from scientists unable or unwilling to spend time straddling the gulf. But he does *not* regard such an exercise as impossible in principle, and it plays a role during conversions after a paradigm shift. Kuhn's view in "Postscript" is thus more moderate than what is indicated by a natural reading of *Structure*, and close to what Quine writes about translations: to say that something is radically different is "to say no more than that the translations do not come smoothly".[43]

How should we understand Kuhn's metaphor that compares the translation of paradigms to the translation of fiction? Translations of fiction include changes of connotations in the stylistic sense; but difference in style would have no significance for our argument. Are there deeper differences? Ian Hacking differentiates two possible positions on translatability problems:

> Quine urges that there is *too much* possibility for translation. The opposed doctrine maintains that there is *too little*. Two human languages could be so disparate that no

[41] P.201.
[42] In Kuhn [1990], p.300.
[43] "Speaking of Objects", in Quine [1969], p.1. There is no reference to Quine in the main text of *Structure*, only in the Preface and "Postscript".

> system of translation is possible. That is in the spirit of Feyerabend's doctrine of incommensurability.[44]

The Kuhn quote about the translation of fiction above indicates that he has a foot in both camps. Multiple translations are possible, but none is *exactly* right.

Wes Sharrock and Rupert Read suggest a way to construe Kuhn's argument that is neither absurd nor contradictory.[45] The first step is to see that in Kuhn's view (by contrast, for example, with Donald Davidson's view[46]), *understanding* is independent of *translating*. For Kuhn, it is possible to understand and describe terms in the conceptual scheme of another paradigm, and, according to "Postscript", the historian of science is in the business, or *should* be in the business, of doing just that.[47]

The second step is to recognise that Kuhn insists that key terms of a scientific scheme often are defined in terms of each other. This makes it impossible, even for someone who understands both schemes, to translate one to the other on a term-by-term or sentence-by-sentence basis; the schemes need to be treated as a whole.[48] *This* is to what incommensurability of translation applies, according to Sharrock and Read.[49]

This sense of 'incommensurability' is plausible, and it complements the description of conceptual changes in scientific revolutions. Concepts change, but not one-by-one; they change as a network of inter-definitions. The "things" that scientific schemes postulate ("the furniture of the universe"), are postulated by the scheme as a whole, not by individual terms, and the whole is needed to understand the parts.

Furthermore, the translation in question is not "radical" in Quine's sense; it takes place from a historical connection and a perspective of familiarity with both the source and the target

[44] Hacking [1975], p.152.
[45] Sharrock and Read [2002].
[46] Davidson [1984].
[47] See p.201.
[48] Kuhn [1990] gives an elaborate example of the relation between 'mass', 'force' and 'weight' in Newton's theory.
[49] They do *not* say that this exhausts what Kuhn means by 'incommensurability'.

language/theory.[50] Finally, if we read Kuhn's examples, translation is perhaps less impossible in principle than tricky in practice, as there are inter-dependencies between terms.

Importantly, however, even if translation issues do not preclude talking (in some sense) "about the same things", we cannot use a translation to justify a continuity of meaning between paradigms. Cross-referencing takes place within the terms of the latter paradigm, with these terms inter-defined by the whole scheme, with no independent success criteria. Even if we accept Sharrock and Read's plausible interpretation of 'incommensurability', it does not solve the problem of determining in virtue of what translations are successful. Davidson's comment is apt: "Kuhn is brilliant at saying what things were like before the revolution using – *what else?* – our postrevolutionary idiom."[51] In other words, there is no escape from our language. In my context, the real problem is how we can justify that terms in theories on different sides of a paradigm shift refer to the same thing – question (A) – in the absence of a neutral language with objective criteria. Kuhn's Challenge remains.

1.8 Other Worlds

We saw that Kuhn claims that proponents of competing paradigms work in different worlds, and that he regards this as the most fundamental aspect of the incommensurability. Kuhn does not in that context go as far as to argue that the real world has changed, or lacks independent existence. On the contrary, he says about two scientists supporting different paradigms: "Both are looking at the world, and what they look at has not changed. But in some areas they see different things, and they see them in different relations".[52] The "different world" picture is compared to a Gestalt

[50] Quine actually dismisses this type of example at an early stage in Quine [1964], p.28: "Translation between kindred languages, e.g., Frisian and English, is aided by resemblance of cognate word forms. Translation between unrelated languages, e.g., Hungarian and English, may be aided by traditional equations that have evolved in step with a shared culture. What is relevant rather to our purposes is *radical* translation, i.e., translation of the language of a hitherto untouched people."
[51] Davidson [1973-1974], p.6. Italics added.
[52] P.149.

switch,[53] and used to explain that paradigm changes cannot be gradual from a psychological point of view: either the scientist sees it one way, or he sees it the other way.

So far, "living in another world" seems to be a metaphor for the barrier scientists have to overcome to be able to look at the world from the perspective of the competing paradigm, contributing heavily to the communication issues that Kuhn means often exist between two paradigms. This is not implausible.

But elsewhere in *Structure* – in Section X – Kuhn takes another step by claiming that paradigm changes also affect the world. At the end of the previous section he writes: "I have so far argued only that paradigms are constitutive of science. Now I wish to display a sense in which they are constitutive of nature as well."[54] His argument for taking this extra step builds on the psychological barriers of scientists just discussed, which make scientific communities look at the world with different eyes. To this he adds a strong metaphysical thesis: "In so far as their only recourse to that world is what they see and do, we may want to say that after a revolution scientists are responding to a different world."[55] We are also told that "the historian of science may be tempted to exclaim that when paradigms change, the world itself changes with them."[56] Again, the absence of communication implied by these different locutions undermines the meaningfulness of talk about scientific progress and implies a negative answer to questions (A) and (B). If scientists before and after a paradigm shift live in different worlds, their theories refer to different objects, and we cannot achieve progress by improved descriptions of the same objects.

Putting to one side what historians of science may or may not be tempted to exclaim, in what sense are paradigms supposed to be "constitutive of nature"? The implication is an idealistic metaphysic that is strongly counter-intuitive. It is also in direct contradiction with the passage I

[53] This analogy is not altogether suitable, as Kuhn later came to realise, as he is describing the conversion of a scientific community, not an individual psychological event. In later texts, Kuhn searches for other metaphors. See Sharrock and Read [2002].
[54] P.110.
[55] P.111.
[56] P.111.

quoted earlier, where Kuhn describes two scientists looking at the world from each side of a paradigm shift: "Both are looking at the world, and what they look at has not changed."[57] Kuhn advances no additional argument in favour of the stronger, metaphysical thesis.

Many have criticised this (hesitant, but still) step into relativism. As we have seen, Davidson points out that Kuhn describes the history of science "using – what else? – our postrevolutionary idiom"[58] and goes on to suggest: "Instead of living in different worlds, Kuhn's scientists may, like those who need Webster's dictionary, be only words apart."[59] Indeed, the problem of identifying paradigms, let alone describing them or understanding them, might well look formidable, if it is to be done from inside another, incommensurable, paradigm. But perhaps the Section X position is not Kuhn's considered opinion. Sharrock and Read comment that these paragraphs "might seem to licence criticism for falling for [relativism]"[60] but that "taken in relation to almost everything else he says, this cannot be right".[61]

Looking for what *is* right, aiming to represent Kuhn's real views under a charitable interpretation, I have already said that "living in another world" could be a metaphor for psychological barriers. The acceptance of a paradigm comes with more than equations and methodologies, it comes with the acceptance of an ontology (a "world view", Kuhn calls it) containing the postulates of the theory. Kuhn also holds that one can only subscribe to one paradigm at the time, creating an element of isolation in the worldview of the paradigm. One interpretation, then, of the "sense" in which the paradigms are constitutive of nature is as a psychological/sociological phenomenon – constitutive of the way the nature *seems to be* for the scientists. This interpretation would obviously not be a threat for conceptual continuity. But it is also clear that this psychological/sociological observation is not everything Kuhn wants to say in Section X.

[57] P.149.
[58] Davidson [1973-1974], p.6.
[59] Davidson [1973-1974], p.11.
[60] Sharrock and Read [2002], p.174.
[61] Sharrock and Read [2002], p.174.

While this is not explicit in the text, I suggest that Kuhn is searching for an interpretation of 'incommensurability' that is *incompatible with realism*. Why does he feel the need to oppose realism? The reason could be: because he sees realism as committed to the search for the one true, perfect theory, towards which science should aim to progress. Kuhn, rightly, regards *that* view as incompatible with scientific practices. I will discuss the notion of the perfect theory in Chapter 5 and identify a way to defend another type of realism. But to do that, I find, we need a solution to Kuhn's Challenge, to which I return in Chapters 6 and 7.

1.9 Theory Comparisons

I earlier formulated two related worries about scientific progress that I called Kuhn's Challenge: *(A) If the meanings of key concepts change between paradigms, how can we say that they are theories about the same thing? And (B) Even if we assume that two paradigms do address the same subject matter, how can we determine which one is better?* I have discussed whether Kuhn's notion of incommensurability excludes the possibility that different paradigms are about the same subject matter. Although the most plausible interpretation of 'incommensurability' does not *exclude* shared subject matter, it does not resolve the problem of how we *justify* translations between terms across paradigms. Furthermore, question (B) requires a meaningful way to compare theories in practice. Kuhn argues that several methods *cannot* be used to compare theories:

> (i) *Absolute comparison.* There is never any inventory over all possible theories out of which the best is chosen, as the testing of competing theories for paradigm-hood focuses only on the choice at hand. "Verification is like natural selection: it picks out the most viable among the actual alternatives in a particular historical situation. Whether that choice is the best that could have been made if still other alternatives had been available…is not a question that can usefully be asked."[62]
>
> (ii) *Fit to data.* There are several reasons this criterion will not work:

[62] P.145.

(ii.a) For a given set of data, there can be more than one theory fitting this set.[63]

(ii.b) Fit to data is always approximate and incomplete.[64] Normal-science activities are needed to make data and theory agree, "to beat nature into line".[65]

(ii.c) Data available for testing are always incomplete and "no theory can ever be exposed to all possible relevant tests".[66]

(ii.d) The scope of paradigms often varies, affecting the question and answers applicable, so that the scope of relevant data is not identical.[67]

(iii) *Absolute progress.* There is no straight line which could be used to locate theories. Kuhn touches on this point already in *Structure*, in his discussion about reversals. After Newton, Kuhn writes, the attempt to explain gravity was "fruitfully abandoned",[68] but that change was reversed by Einstein, who provided explanations that were "in this particular aspect, more like those of Newton's predecessors than of his successors."[69] Looking at Aristotle, Newton and Einstein, Kuhn adds (in "Postscript"): "I can see in their succession no coherent direction of ontological development."[70]

(iv) *Closeness to truth.* Regarding the notion of truth, Kuhn is sceptical. Towards the end of *Structure*, he notes that he has not so far used the notion at all in the book, with the exception of a quote from Francis Bacon. At this point, he questions the need to have truth as a teleological goal, towards which science progresses. The philosophy of science does not need goals any more than the theory of evolution does, he says.[71]

[63] P.76.
[64] P.146.
[65] P.134.
[66] P.144.
[67] P.109.
[68] P.108.
[69] P.108.
[70] P.205.
[71] P.169, p.205.

(v) *Comparison using neutral standards.* "[T]here can be no scientifically or empirically neutral system of language or concepts…[T]ests and theories must proceed from within one or another paradigm-based tradition."[72] and "There is no neutral algorithm of theory-choice, no systematic decision procedure which, properly applied, must lead each individual in the group to the same decision."[73]

Now, it is not Kuhn's view that theories cannot be compared at all. However, his account of how we compare does not answer (A) and (B).

Kuhn accepts that the question of the quality of a paradigm is meaningful, not in comparison with an ideal, true theory that we can never have, but in comparison to another, competing theory. It makes no sense to ask in absolute terms whether a theory agrees with the facts, because: "All historically significant theories have agreed with the facts, but only more or less."[74] However, "It makes a great deal of sense to ask which of two actual and competing theories fits the facts *better*."[75] But having said that, he soon complicates the picture: "This formulation, however, makes the task of choosing between paradigms look both easier and more familiar than it is, because it depends on conditions never met completely"[76] due to the meaning changes that come with the new paradigm. He concludes: "The competition between paradigms is not the sort of battle that can be resolved by proofs."[77] We can meaningfully *ask*, it appears – but the question is what kind of answer can be given. I will discuss this in Chapter 6.

In "Postscript", rejecting accusations of relativism, Kuhn lists some criteria used to detect progress in an evolutionary tree over scientific theories from primitive, common beginnings to modern theories. Lines drawn up that tree from the trunk to the tip of some branch would "trace a succession of theories related by descent."[78] At least for theories not too far apart, Kuhn says, it

[72] P.145.
[73] P.198.
[74] P.146.
[75] P.146.
[76] P.147.
[77] P.147.
[78] P.204.

should be easy draw up criteria to distinguish the theories and to tell earlier from later stages of development. He suggests suitable candidates.

> Among the most the most useful would be: accuracy of prediction, particularly quantitative prediction; the balance between esoteric and everyday subject matter; and the number of different problems solved. Less useful for this purpose, though also important determinants of scientific life, would be such values as simplicity, scope, and compatibility with other specialities.[79]

Kuhn does not state that this is the definite list, but he is optimistic: "Those lists are not yet the ones required, but I have no doubt that they can be completed."[80] When they *are* completed, Kuhn writes, scientific development could be described as "a unidirectional and irreversible process."[81]

The "Postscript" comparison parameters take the evolutionary tree of theories as their starting point. But what is the criterion for putting two theories in the same part of the tree? A comparison using the criteria Kuhn lists will only work for two theories having the same subject matter in some sense; explanatory power, simplicity, and so on, cannot be absolute. Whether Darwin's theory of evolution is better or worse at predictions than Einstein's general theory of relativity is not normally an interesting question. The tree metaphor addresses this issue and suggests a historian's perspective; one theory *in fact* followed another. Einstein's theory followed Newton's, not Darwin's. But to avoid relativism, this does not seem to be enough, without further justification; we want to say that there is a sense in which Einstein and Newton have very significant overlaps in terms of addressing the same subject matter, the same phenomena. Question (B) can only be addressed in conjunction with question (A); it presupposes an answer to (A). Successful comparison between theories in the sense intended in (B) assumes that we are comparing theories *about the same things*.

[79] P.204.
[80] PP.204-205.
[81] P.205.

To postulate criteria for how theories can be compared therefore reintroduces the issue with question (A). These criteria either work from the perspective of a later paradigm ("victors writing history"), or assume a constancy of subject matter across paradigms that we are trying to justify.

1.10 Conclusions

We firmly believe that our mature sciences have made enormous progress, and we want to have a philosophical model to justify this belief. The detailed and insightful description of scientific practices in *Structure* describes the issue at hand, resulting in what I have called "Kuhn's Challenge":

> (A) *If the meanings of key terms change between theories on either side of a paradigm shift, how can we say that these theories are about the same thing?* And,
> (B) *Even if we assume that two theories do address the same subject matter, how can we determine which one is better?*

Kuhn's description of paradigm changes, underpinned by his notion of incommensurability, gives a plausible and informative description of the relationship between groups of scientists around a paradigm change, what I have called "Kuhn's sociological thesis". But applied to the relationship between the paradigm themselves, incommensurability is more problematic.

Kuhn *could* give up on questions (A) and (B) altogether and adopt a relativist position, but this is something he strongly rejects in "Postscript". And would he have chosen to adopt this position, he would have held a view that denies scientific progress, which I exclude by fiat from the scope of this thesis. My discussion is about philosophical models of scientific progress, and the *explanandum* is actual progress. The relativist option does not so much stop us in our tracks as take us off the rails altogether. I have not found anything in the nature of incommensurability motivating such a drastic conclusion. Changes are real – but cross-referencing is possible. Scientists talk at cross-purposes – but historians can act as translators. To convert to a new paradigm might feel like moving to a different world – but it isn't really.

However, even if incommensurability as a threat to the justification of scientific progress can be largely disarmed, Kuhn's Challenge still remains. How do we justify scientific progress in the light of conceptual change?

Does Kuhn believe that there are satisfactory responses to (A) and (B)? At least in "Postscript", he clearly believes that there are good answers to (B), so he *should* also hold that there are good answers to (A), as (B) depends on (A). But there is no elaborated, positive answer to (A) in *Structure*; we learn more about what does *not* count as an adequate response. However, there are elements in *Structure* that indicate one type of answer, and I will come back to this in Chapter 7.

After Kuhn, one popular response to Kuhn's Challenge has been to stress the role of reference rather than meaning for continuity, and to rely on the semantic/modal arguments put forward by Saul Kripke and Hilary Putnam. I will introduce and analyse these arguments in the following chapters, but conclude that on their own, they do not suffice for this purpose.

Many writers have also developed ideas similar to those found in *Structure*, and I will later in this thesis discuss Hasok Chang, John Dupré, Michael Friedman, Muhammad Ali Khalidi, Joseph LaPorte and P D Magnus. With their help, I will construct a framework, largely compatible with the Kripke-Putnam semantics, that can provide a better response to Kuhn's Challenge.

2 Kripke's Semantics

2.1 Introduction

Some accept the threatening relativism implied by Kuhn's Challenge[82] because they embrace incommensurability. Among those who instead try to find counter-arguments against Kuhn, some want to utilise Saul Kripke's and Hilary Putnam's semantical apparatus. This includes Putnam himself in "The Meaning of 'Meaning'",[83] as we will see in the next chapter. Dudley Shapere calls this "the most influential approach to the incommensurability claim."[84]

One reason this type of response looks promising for the purpose is that it separates what had traditionally been called "meaning" from reference. If the former is affected by theory-changes, maybe the latter, thus liberated, might not be. Earlier speakers may have *meant* something rather different when talking about (for example) planets, but if there is a chain of reference that is independent of descriptions, that guarantees continuity; despite paradigm changes, we can still be talking about the same things. In this way a theory built on the Kripke-Putnam semantics[85] can recognise Kuhn's arguments about paradigm shifts, but hold that this is less important than Kuhn thought. Thus, the Copernican Revolution, according to Bird, "does not establish any shift of extension rather than a shift in what the extension was thought to be."[86] Howard Sankey describes the intended conclusion as follows:

> But if reference is stable through conceptual change, the terms employed by alternative scientific theories may share reference despite variation in concepts. Given shared reference, statements from meaning variant theories may enter into conflict or agreement,

[82] Kuhn himself did not accept this conclusion, but continued to struggle with the incommensurability concept in later articles, see Kuhn [1990] and Kuhn [2002].
[83] Putnam [1975], chapter 12.
[84] Shapere [1998], p.735.
[85] The term "Kripke-Putnam" semantics can be objected to: while there are many important similarities, there are also differences between the two philosophers, at least in emphasis and focus, and I will introduce them separately. Ian Hacking [2007] discuss the differences between Kripke's and Putnam's view of natural kind terms.
[86] Bird [2004], p.53.

since their component terms refer to the same things. Hence such theories may be compared for content, and are not therefore incommensurable.[87]

In principle, this would answer both question (A) *("If the meanings of key terms change between theories on either side of a paradigm shift, how can we say that these theories are about the same thing?")* and question (B) ("*Even if we assume that two theories do address the same subject matter, how can we determine which one is better?").* I will argue that it is doubtful whether Kripke himself has the ambition to answer Kuhn's Challenge. However, two types of arguments emanating from his theory have been used by other philosophers to construct such an answer. I will present their background in Kripke's and Putnam's machinery before I turn to the arguments themselves.

In his main work, *Naming and Necessity* (*N&N*),[88] Kripke pays most attention to proper names, criticising the received view at the time. This is where Kripke establishes his analysis, where it is most detailed, and where it has met the widest acceptance. In a second step, Kripke extends his arguments to so-called "natural kind terms"; examples include 'gold', 'heat' and 'tigers'. He argues that the semantics of these terms is very similar to that of proper names. I will follow the same order and start with proper names, first with Kripke's negative and then his positive arguments, before I proceed to natural kind terms and scientific identifications. In the next chapter I will continue the discussion of natural kind terms, mainly taking examples from Putnam instead. I will identify some assumptions underpinning Putnam's arguments, relevant to address Kuhn Challenge. In later chapters I will analyse and question those assumptions. But in this chapter, I will claim that Kripke's own semantics is not intended as a response to Kuhn.

[87] Sankey [1997], p.429.
[88] Kripke [1981].

2.2 Descriptionism for Proper Names

The primary target for Kripke in *N&N* is the theory he calls "descriptionism",[89] which relies on descriptions of unique properties to explain the meaning and reference of terms. Many of Kripke's ideas in *N&N* are developed in his criticism of descriptionism.

This theory is associated with Gottlob Frege[90] and Bertrand Russell,[91] but also defended by more recent philosophers such as John Searle.[92] I am interested in this theory as far as it serves as a target for Kripke and will not be overly worried about whether he does them full justice, or with the (significant) differences between these philosophers.[93]

The descriptionist theory has a number of applications. Applied to proper names, it can be seen as a response to the common-sensical theory of names advocated by John Stuart Mill.[94] Mill defended the view that "singular terms" (including proper names) lack meaning, as opposed to "general terms" (including natural kind terms) which do have meaning – his word is "connotation". The only semantic content of a proper name, according to Mill, is its reference – its "denotation".

Mill's theory got into trouble. One issue is names that lack a referent, but where there nevertheless appears to be some semantic content: there is a difference between 'Harry Potter' and 'Lady Macbeth'.

Another major problem is posed by identity statements such as 'Hesperus = Phosphorus',[95] which expresses the discovery that 'Hesperus' and 'Phosphorus' do not pick out different stars, but instead are both names for the planet Venus. 'Hesperus' and 'Phosphorus' have something in common, the common reference, but also something that separates them. It is perfectly possible to believe something about Hesperus without believing the same thing about Phosphorus, and

[89] Elsewhere sometimes called "descriptivism".
[90] Frege [1980].
[91] Russell [1905] and [1919].
[92] E.g. Searle [1958].
[93] For a recent discussion comparing Frege, Russell and Kripke, see Colin McGinn [2015].
[94] Mill [1973-1974].
[95] This is Russell's version – Frege talked about "the Morning Star" and "the Evening Star".

indeed to believe that the two are not identical, without being conceptually confused, just badly informed about our planetary system. Thus a related puzzle arises concerning belief attributions: it is possible that 'Paul believes that Hesperus is a planet' can be true while 'Paul believes that Phosphorus is a planet' is false. The terms seem to convey different information, and to contribute something different (a different "semantic content") to the truth conditions of the sentences above.

Frege addresses these perceived shortcomings in Mill's theory by introducing another component, similar but not identical to Mill's notion of connotation, namely *Sinn* (sense). 'Hesperus' and 'Phosphorus' have the same reference, Venus, but different Fregean senses, and this makes the difference in their contribution to the meaning of sentences like those about Paul above. A Fregean sense is a technical notion that Frege leaves relatively vague, but which is a "mode of presentation" of the referent. In a footnote to his paper "On Sense and Reference",[96] Frege mentions an example, namely that that the sense of 'Aristotle' could be the description 'The pupil of Plato and teacher of Alexander the Great'.[97] This is a description that points to exactly one man, Aristotle. Sense is thus often taken to be a *descriptive* mode of presentation, functioning similarly to definite descriptions in Russell's theory of names. Because this is how Kripke interprets the notion in *N&N*, I will understand 'sense' in this way here.

Descriptionism (at least according to Kripke) also holds that to master a term, that is, for example, to be a competent user of a proper name, we need to know the Fregean sense of that term. I understand what 'Aristotle' means if and only if I know that he was the pupil of Plato and the teacher of Alexander the Great.

If this is right, it appears that the statement 'Aristotle is the pupil of Plato and the teacher of Alexander the Great' is *necessarily* true in descriptionist semantics, because there are no possible situations where we could correctly speak about Aristotle without speaking about the pupil of Plato and teacher of Alexander the Great (assuming that sense is shared in the community). It is not possible to *be* Aristotle without being the pupil of Plato and the teacher Alexander. It is also

[96] Frege [1960].
[97] Frege [1960], footnote 4. This idea is developed in Russell [1905] and [1919].

true *a priori*, as it is true in virtue of the meaning of the terms: any speaker who understands 'Aristotle' also understands that it means 'the pupil of Plato and the teacher of Alexander the Great', and knows that 'Aristotle is the pupil of Plato and the teacher of Alexander the Great' is true in virtue of this understanding alone. For descriptionist semantics, *apriority* and necessity coincide.

For descriptionism, the Fregean sense is different from the reference, but the sense *determines* the reference. In Frege's example, the sense of 'Aristotle', *being the pupil of Plato and the teacher of Alexander the Great*, determines that 'Aristotle' refers to Aristotle himself and nobody else. To talk about Aristotle is to talk about the man who has this property. The Fregean sense gives us an explanation of what it is to master a proper name, what defines its reference, and how a proper name contributes to the truth conditions of a sentence.

To sum up, the core of descriptionism for proper names, at least as Kripke understands it in *N&N*, is the following:

a) A description of uniquely identifying properties is the Fregean sense of a proper name.
b) The Fregean sense is the semantic content (the meaning) of a proper name, what it contributes to the truth conditions of a sentence containing it.
c) To understand and be able to competently use a proper name, we must grasp its Fregean sense.
d) The Fregean sense determines who/what the name refers to.

On top of this, the theory also implies:

e) The sense of a proper name is known *a priori*.
f) It is a necessary truth, if it is a truth, that a proper name has a certain sense.
g) If the sense does not identify a (unique) referent, the proper name does not refer to anything.

2.3 Kripke's Criticism of Descriptionism

Descriptionism is an elegant and powerful theory, and Kripke pays his tribute. It has only one disadvantage, he says: it disagrees with the facts. To show this, Kripke uses a series of philosophical thought experiments. In one type of thought experiment, he varies our epistemological situation: he imagines that we find out something we did not know before. In another, he puts us in a counterfactual situation, which in some specified way differs from the real world.[98]

Kripke's fundamental objection to descriptivism targets what might look like a strength: the double duty for the Fregean sense. The senses both make it possible to understand what a term means and determine the reference. He writes:

> Frege should be criticized for using the term 'sense' in two senses. For he takes the sense of a designator to be its meaning; and he also takes it to be the way its reference is determined…They should carefully be distinguished.[99]

Kripke subsequently criticises the theory both as a theory of meaning and as a theory of reference.

Descriptionism is not an adequate theory of meaning, he states, because what we mean by 'Aristotle' is not 'the pupil of Plato and the teacher of Alexander the Great'. To say that Aristotle was the pupil of Plato is not a tautology, and if we found out that he was never taught by Plato, we would not conclude that Aristotle never existed. "Most of the things commonly attributed to Aristotle are things that Aristotle might not have done at all. In a situation in which he didn't do them, we would describe that as a situation in which *Aristotle* didn't do them."[100] Descriptions of the properties of Aristotle that Kripke has in mind here are all at best *contingently*, not necessarily, true of Aristotle, as they describe properties he might have lacked. Some might even

[98] I will discuss the issue of where these facts come from in Chapter 3.
[99] Kripke [1981], p.59.
[100] Kripke [1981], p.61.

be false. Such descriptions therefore cannot be what the name means, or as Kripke also puts it, they cannot be *synonyms* of the name.

Searle has in his version of descriptivism introduced the term "cluster of descriptions", which is a *disjunction* of descriptions, not all of which have to be true for reference to be successful. We could add more descriptions to this cluster. We could for instance add that Aristotle was *fond of dogs* or *born in Stagira*. Searle's version avoids some counterexamples that were problematic for the original version, as he does not need to say that a particular description is synonymous with the proper name, nor that it is fatal if one or two turn out to be false. Maybe a majority would have to be true, and maybe there is a weighting. But Kripke says that Searle's idea still falls short. If we found out that we were mistaken also regarding the preference for dogs due to an early historian mixing up the great philosopher with an obscure local dog lover, this would still not lead us to conclude that Aristotle never existed. The same is true for the place of birth. We could be mistaken on each point. Adding further properties to the disjunction does not seem to help; we could find that we were really thoroughly mistaken and that Aristotle actually lacked all of them, including being called "Aristotle", without being forced to say that there was no Aristotle.[101] We can always be mistaken in the descriptions we associate with a name. Consequently, the cluster of properties cannot be what a proper name means; it cannot act as a synonym.

> To me Aristotle's most important properties consist in his philosophical work, and Hitler's in his murderous political role; both…might have lacked these properties altogether. Surely there was no logical fate hanging over either Aristotle or Hitler which made it in any sense inevitable that they should have possessed the properties we regard as important to them; they could have had careers completely different from their actual ones.[102]

[101] With one qualification, which I will come back to: if the property non-typically happens to be a necessary one.
[102] Kripke [1981], p.77.

It is useful here to introduce a term Putnam coins: he calls descriptions of properties "commonly attributed" to objects their "stereotypes".[103] Searle's cluster of descriptions would constitute the stereotype associated with 'Aristotle'. Stereotypes are relevant to linguistic proficiency, to master the term in normal usage. But they are not Fregean senses, as they do not contribute semantic content to sentences featuring proper names. We cannot substitute a name for its stereotype in a statement without changing the meaning of that statement.

Even if descriptionism does not work as a theory of meaning, it could in theory still work as a theory of reference. But in normal circumstances, the same examples show that stereotypes do not determine references either.[104] Even if Aristotle in fact *did* have some or all of the properties assigned to him by the stereotype, we can easily imagine a counterfactual situation where he did not. And even if we could identify some properties such that Aristotle must have them to be Aristotle, properties that are necessary and perhaps sufficient conditions for being Aristotle, we do not need them to refer successfully. As Kripke says about Nixon:

> [E]ven if there were a purely qualitative set of necessary and sufficient conditions for being Nixon, the view I advocate would not demand that we find these conditions *before* we can ask whether Nixon might have won the election…[105]

To summarise: A cluster of descriptions, a stereotype, is not the semantic content of a proper name and it does not determine its reference. A description of contingent properties such as 'the pupil of Plato and the teacher of Alexander' could replace the name 'Aristotle' in some ordinary discourse contexts, but the difference shows up in the thought experiments that vary our metaphysical and epistemological situations. Proper names may be associated with stereotypes, but in Kripke's opinion, they lack a Fregean sense. The stereotype cannot be the semantic

[103] In Putnam [1975], chapter 12. Kripke himself moves in this direction after the *N&N* lectures. In Kripke [1981], p.163, point **(e)** of the Addenda, he refers to "the predominantly social character of the use of proper names". In Kripke [1973], p.65, Kripke makes use of Putnam's notion directly: "If one is referring to an actual animal, one may of course pick it out by what Putnam calls a 'stereotype'…without knowing what its internal structure is". See Fernández Moreno [2016] for discussion.
[104] In Kripke's opinion, descriptions *do* determine references – but only rarely, in specific situations that I will discuss in §2.5.
[105] Kripke [1981], p.47.

content of the name because it is not synonymous with the name. And the stereotype cannot (normally) determine reference either, as it could be wrong, or only contingently true. Kripke draws the conclusion that the only semantic content of a proper name is the referent. This puts him close to Mill's position, which Kripke acknowledges: "My own view…regards Mill as more-or-less right about 'singular' names".[106]

2.4 Rigidity and Necessity

Importantly for the later application to natural kind terms, Kripke's criticism of descriptionism assumes that there is no problem with identity across possible worlds ("trans-world identifications"). If we talk about what would have happened to world politics if Nixon had lost the 1968 election he in fact won, we are still, Kripke insists, talking about *exactly the same man*, Richard Nixon, but in a counterfactual situation. This situation is one that did not actually occur, but it *could* have occurred, though Nixon would then have had different properties than he had in the real world (those related to election success). The reason that trans-world identifications are unproblematic for Kripke is that proper names are what he calls "rigid designators". 'Designators' stands for names or descriptions, and 'rigid' means that the designator refers to the same object in all possible situations – in all possible worlds.[107] Nixon having lost the election is still Nixon. A clone of Nixon, however similar, is not him.

> [I]t is *because* we can refer (rigidly) to Nixon, and stipulate that we are speaking of what might have happened to *him* (under certain circumstances) that 'transworld identifications' are unproblematic in such cases.[108]

Rigidity is powerful, due to its connection with necessary truths: true identity statements between rigid designators are always necessarily true.[109] For proper names, this appears to hold in virtue of only the standard identity relation and the two rigid designators. Because rigidity is so powerful and so central to Kripke's argument, he needs a method to find rigid designators, and

[106] Kripke [1981], p.127.
[107] See Kripke [1981], p.48. I ignore for the moment the complication of worlds where the object does not exist. But this will become an issue when rigidity is applied to natural kind terms.
[108] Kripke [1981], p.49.
[109] But not vice versa: there can be necessarily true identities between non-rigid designators.

he needs success criteria. It is not enough to look at what is the case in the actual world, because rigidity bridges our world with other possible worlds. Kripke's method to identify rigid designators is the thought experiment, and the success criteria our linguistic intuitions.[110]

For completeness and for later reference,[111] I will need to mention another important case. I said that according to Kripke we can successfully refer with a description that an object might have lacked or in fact did lack. But if the property used for reference *is* necessary for an object, the corresponding description is necessarily true, and we cannot be mistaken in the same way that we can be regarding a contingent property. We cannot find out that it lacks the property in question, because an object without that property would not be the same object. But we could be in an *epistemologically* indiscernible situation. For example, Kripke believes that specific biological origin is necessary for being a certain person. If that is right, we could *not* find out that Aristotle had different parents to the ones he in fact did, but we *could* imagine confusing Aristotle with a Doppelgänger who had different parents.

For descriptionism as Kripke defines it, all necessary truths are also *a priori* truths. But it is an important part of Kripke's theory that necessary properties, such as the biological origin of Aristotle, can be unknown. That some statements are necessarily true, does not imply that they are known *a priori*. What Kripke means by *a priori* and *a posteriori* is not totally clear in *N&N*, and he indeed cautions against using the term '*a priori*' at all, except to characterise the epistemic situation of individuals.[112,113] But the reason why these notions can remain at an intuitive level is that what is most important for Kripke in *N&N* is not the nature of either *apriority* or necessity, but that *apriority* is not typically coextensive with necessity. When

[110] My main discussion of thought experiments is in §§3.3-3.5.
[111] This will become important in chapter 8.
[112] Kripke [1981], p.35: "It might be best therefore, instead of using the phrase *'a priori* truth', to the extent that one uses it at all, to stick to the question of whether a particular person or knower knows something *a priori* or believes it true on the basis of *a priori* evidence".
[113] He does however not always follow his own advice; see e.g. his definition of an analytic truth. Kripke [1981], p.122, note 63, says that this "depends on *meanings* in the strict sense and therefore is necessary as well as *a priori.*"

transferred to natural kind terms, this distinction gives us a model for how discoveries are possible, and it is a cornerstone for his analysis of scientific identifications.

2.5 Kripke's Positive Theory of Proper Names

N&N does not include a comprehensive, positive theory about the meaning of proper names (nor the meaning of natural kind terms). This has sometimes been regarded as an omission to be corrected. Linsky writes:

> While Kripke did give an account of the reference of names, he did not give a positive account of the *meaning* of a name. Numerous philosophers have proposed to supplement Kripke…by following Frege…[114]

The lack of such a theory is arguably natural, as Kripke is of the opinion that proper names do not have meaning in the traditional sense. As we saw, Kripke's references to Mill would back up that interpretation. However, it is better taken with some caution. We should be careful not to say that Kripke's view is that descriptionism is wrong in all respects or that descriptions have no role in the semantics of proper names.

According to Kripke, there are several ways in which descriptionism is wrong, as we have seen. But there are also features of descriptionism that Kripke agrees with, although this tends not to be spelt out in *N&N*. For example, in some cases, Kripke thinks, the theory *does* provide a good description of reference-fixing. These are situations when an object is given a name by definition in a literal or metaphorical baptism: "I will call the species these bones belong to 'Tyrannosaurus Rex'", or "I define one meter as the length of this bar".[115] But this is not the way most speakers refer to the dinosaur or the measurement. Most speakers use the terms in the same way as they have been used before by other speakers, without direct links to the bones or the iron bar. Here Kripke outlines an alternative theory of reference both for proper names and for natural kind terms.

[114] Linsky [2011], p.20.
[115] Descriptions can either be used to define the meaning or to fix the reference. See Kripke [1981], p.59.

According to Kripke, we can continue to refer to the same object after the initial baptism by a causal chain of usage that will help to secure successful continuity of reference.

> An initial 'baptism' takes place. Here the object may be named by ostension, or the reference of the name may be fixed by a description. When the name is 'passed from link to link' the receiver of the name must, I think, intend when he learns it to use it with the same reference as the man from whom he heard it.[116]

The initial event, then, might involve pointing, but it also might involve a description. This is an area where Kripke's view seems to have evolved. In his article "Identity and Necessity" ("I&N") he writes: "[T]he reference of names is rarely or almost never fixed by means of description."[117] In the main text of *N&N*, he instead states: "In an initial baptism [the reference] is typically fixed by an ostension or a description."[118] In a footnote to *N&N*, he expresses a slightly different opinion again: "The case of a baptism by ostension can perhaps be subsumed under the description concept also. Thus the primary applicability of the description theory is to cases of initial baptism."[119,120] But a central point for Kripke remains unchanged, namely that there can be successful reference-fixing using descriptions of contingent properties. What is also unchanged is that while the baptism might be based on a description, but subsequent referencing relies only on the speaker's intention to use a term in the same way as previous speakers.[121] Kripke repeatedly stresses that he is not presenting a complete theory, only "a better picture than that given by description theorists."[122]

I said that Kripke believes that the reference of a proper name is fixed to a person or an object in an initial baptism. But importantly, the descriptions used in reference-fixing are not necessarily true, they do not *determine* the reference. Reference-determination is a matter for the properties

[116] Kripke [1981], p.96.
[117] Kripke [1972], p.157.
[118] Kripke [1981], p.135.
[119] Kripke [1981], p.96 note 42.
[120] I will discuss certain problems with the notion of reference passing "from link to link" in the next chapter, together with Putnam's more elaborated solution.
[121] Kripke [1981], p.96.
[122] Kripke [1981], p.97.

that make an object into the object it is: its *essential* properties.[123] With that in mind, I can now explain my earlier claim that descriptions do not *normally* determine the reference, as they can be wrong or contingent. Descriptions of essential properties, however, are necessarily true, and they do determine the reference.[124]

I will discuss essences in much more detail in connection with natural kinds, starting in the next section.

2.6 Theoretical Identifications and Natural Kind Terms

Kripke largely follows Mill in his view on proper names. But he differs from him on natural kind terms; for Mill, those and other general terms have connotations. In contrast, Kripke in the third *N&N* lecture states that "terms for natural kinds are much closer to proper names than is ordinarily supposed"[125] and that they "have a greater kinship with proper names than is generally realized."[126] The natural kind terms he has in mind are vernacular, everyday terms like 'water' or 'gold'.

There have been different opinions about the right interpretation of this "greater kinship". The relatively sketchy treatment in *N&N* makes Scott Soames regard it as a part of an "unfinished semantic agenda",[127] and both Soames and others sympathetic to Kripke's theory have tried to supply the missing piece of semantics. Others, as we will see, have argued that these missing pieces cannot be found.

In this section, I will outline the ways I believe that Kripke thinks natural kind terms are similar to proper names, and explore them one by one. The main ones are:

---

[123] Kripke [1981], p.48: "When we think of a property as essential to an object we usually mean that it is true of that object in any case where it would have existed." And on p.53: "Some properties of an object may be essential to it, in that it could not have failed to have them." But (p.53) these essential properties of an object need not be "the properties used to identify it in the actual world".

[124] In *N&N*, this is described as the exception to the general rule. One example is Kripke [1981], p.60, when he states that Nixon might have lacked any of the properties we associate with him "except that some of these properties may be essential."

[125] Kripke [1981], p.127.

[126] Kripke [1981], p.134.

[127] Soames [2003].

1. Natural kind terms are rigid, just like proper names.
2. Scientific identifications – such as 'water is H_2O',[128] 'heat is the motion of molecules' and 'gold is the element with atomic number 79' – like identity statements with proper names are necessarily true if true at all.
3. In both cases, it is the rigidity that guarantees the necessity of identity statements.
4. In neither case do stereotypes *necessarily* determine the reference or extension, and typically they do not.
5. In both cases is it possible to use contingently true descriptions to fix the reference/extension in a baptism.
6. The reference/extension of both proper names and natural kind terms is at least in principle determined by their essential properties.[129]

As was the case for proper names, Kripke's arguments rely on thought experiments. One example is gold. Kripke imagines a situation where gold turns out not be yellow at all.[130] Maybe we have been the victims of optical illusions or the work of a demon, but we eventually find out that gold is actually always blue. Would we now say that there is no gold? Kripke answers:

> Would there on this basis be an announcement in the newspapers: 'It has turned out that there is no gold. Gold does not exist. What we took to be gold is not in fact gold'...It seems to me that there would be no such announcement. On the contrary, what would be announced would be that though it appeared that gold was yellow, in fact gold has turned out not to be yellow, but blue.[131]

But not all the properties have the same status. It is not Kripke's view that we could find *all* descriptions of gold to be false, but only that we might find out that all descriptions we associate with gold are mistaken. What we could *not* find out, Kripke insists, is that something is gold that does not have atomic number 79 in the periodic table. In a situation where a metal is found with

[128] I will assume throughout that this identification is true, although it arguably is not literally true due to the existence of other chemicals in normal water.
[129] In the next section I will claim that natural kind terms can be seen as having abstract objects as their referents.
[130] To simplify, I will pretend that gold always is yellow in the actual world.
[131] Kripke [1981], p.118.

all the properties of gold, but is constituted of another substance "[o]ne should not say that it would still be gold in this possible world, though gold would then lack the atomic number 79. It would be some other stuff, some other substance."[132]

The gold example supports most of Kripke's points above. Using the colour and other contingent observational criteria, we are able to fix the reference to gold in a baptism (point 5). But such properties do not qualify as being as necessary ones, as Kripke's variation of our epistemological status is designed to show. If we found out that we have been wrong about the colour of gold, the bars, coins and samples we have so far been calling "gold" would continue to be so called. The thought experiment establishes something that is not visible by just looking at the world, where gold is yellow and hard. These observational properties, part of the stereotype, are not the properties that determine the reference. Just as we can refer to a Nixon who lost the election, we can refer to gold of another colour than yellow (point 4). My qualification "*necessarily* determine the reference" is there to cover the case where the stereotype happens to contain a necessary property – but that is not the case here.

In particular, we recall that Kripke's perhaps most significant disagreement with the description theory of proper names is the link this theory makes between assigning meaning and fixing reference. Kripke criticises Fregean senses for mixing up two things that should be separated, namely the meaning of a name on the one side and the reference-fixing on the other. This applies to natural kind terms too.

> [I]n the case of species terms as in that of proper names...one should bear in mind the contrast between the *a priori* but perhaps contingent properties carried with a term, given the way its reference was fixed, and the analytic (and hence necessary) properties a term may carry, given by its meaning. For species, as for proper names, the way the reference of a term is fixed should not be regarded as a synonym for the term.[133]

[132] Kripke [1981], p.118.
[133] Kripke [1981], p.135.

Having argued that the stereotype does not contain the essence of gold, Kripke goes on to say what does: a description of its microstructure. If we find a sample that looks like gold, but has another microstructure, that is, if we find a sample that does not have atomic number 79 in the periodic table, we would not call it "gold" (point 6). This identity between the natural kind and its microstructure is apparent in the epistemic thought experiment. If we construe a modal version of the experiment, we see that it also holds across all possible worlds, and this goes both ways: all gold samples consist of the element with atomic number 79, and all instances of this element are gold. Both 'gold' and 'the element with atomic number 79' are rigid designators, according to Kripke's definition (point 1). And as before, if we accept that they are rigid, it follows that a statement of an identity between them is necessarily true, if true at all (point 2).

There are, however, several challenges to extending the account of proper names to natural kind terms. I consider these in the next few sections.

2.7 The Reference of Natural Kind Terms

The reference of a proper name is well grounded by its relation to an object with a timeline. 'Richard Nixon' refers (if uttered with the right intention) to a person who (in the real world) was the 37th U.S. president, born in 1913 and dead in 1994. What natural kind terms refer to is less obvious.

One natural thought is to regard natural kinds as classes of individuals with some important common properties, and to see these properties as universals with instances. So, if the reference of the natural kind *gold* were the set of all individual gold samples, the common property would be (at least) to have a certain chemical structure.

But this line causes a problem for Kripke's project. If natural kind terms refer to sets of individuals, the fact that one specific individual might not exist in another possible world threatens the stability of reference required for rigidity. I said that 'Richard Nixon' refers to the same individual in all possible worlds, but that needs a qualification since individuals do not

exist with necessity.[134] We can imagine a world where Hannah and Francis Nixon, Richard Nixon's actual parents, have no children. 'Richard Nixon', we should say, refers rigidly to the same individual in all possible worlds *where he exists*. For natural kind terms, this is why we cannot both think of the term 'tiger' as referring to the set of all individual tigers and maintain that 'tiger' refers rigidly, because the set varies across possible worlds.

We cannot solve the issue by an *ad hoc* stipulation that 'tiger' refers rigidly to all *existing* tigers, or even to all tigers who ever have or ever will exist. This is because animals can be subject to scientific re-classifications during their lifetime. Some whales roaming the oceans at the critical time started their lives classified as fish and ended them classified as mammals.[135] Kuhn argues that due to theory and conceptual changes, an initial baptism ("dubbing") is not enough to fix the reference of a term; it needs to be repeated when the changes occur ("redubbing"). From this Kuhn draws a conclusion about rigidity: "Only for the periods between [the redubbing events] does dubbing result in rigid designation."[136,137]

The natural thought that natural kind terms refer to a set of individuals turns out to be incompatible with the idea that they are rigid. But kind terms *are* rigid – this is what Kripke's thought experiments establish. It is a feature of natural kind terms to have a stable reference across time and possible worlds. If so, they cannot designate sets of individuals. A better option might be to look at natural kinds as abstract objects, with properties of their own.[138] This makes Kripke's theory stronger and the correspondence to the proper name situation closer.

Postulating kinds as something over and above, and not reducible to, individuals can be motivated by methodological considerations: it is difficult to see how some scientific statements about kinds can be replaced with statements about individuals. David Armstrong discusses one of Quine's examples from this perspective, namely: "Some zoological species are cross-

[134]Better: it is not a necessary property of any world W that any individual P exists.
[135] Kuhn, in Kuhn [2002], p.205, came to hold that "redistribution of individuals among natural families or kinds… [is]…a central (perhaps *the* central) feature of…scientific revolutions." An opponent could say that the previous classification was incorrect. I will come back to this type of argument in later chapters.
[136] Kuhn [1990], p.298.
[137] Kuhn's criticism of the Kripke-Putnam semantics for natural kind terms is discussed in Kuukkanen [2010].
[138] See Donnellan [1983] and LaPorte [2004].

fertile".[139] And Dupré writes: "For example, ecology, involving the interaction between different *kinds* of organisms, seems to be impossible to formulate in any but the most abstract terms without treating specific kinds of organisms as kinds."[140] Both cases suggest that we naturally assign properties to kinds directly, in a way that is not just an aggregation of the properties of individuals, and that kinds are in this practice implicitly treated as abstract objects. This way, natural kind terms refer to single objects, just like proper names. Keith Donnellan believes that

> construing words for kinds, such as 'water', 'tiger', etc., as rigid designators *and* giving the Kripke-Putnam view the best run for its money is to think of them as what Mill calls 'abstract' nouns…Thought of in this way, kind terms are in one way like proper names: they designate a single entity, albeit an abstract entity – a species or a substance in these cases.[141]

However, if the reference of a natural kind term is to an abstract kind, as opposed to a set of individuals, there is still a problem with rigidity, but the opposite one: it has been argued that rigidity comes suspiciously cheaply. Helen Beebee and Nigel Sabbarton-Leary believe that

> it trivializes the notion of rigidity, since *all* general terms…turn out to be rigid…If we treat general terms as the names for kinds, then just as 'water' rigidly refers to the water-kind, so 'fridge' refers to the refrigerator-kind.[142]

The underlying issue is whether it is the job of the rigidity notion to separate terms standing for true natural kinds from their illegitimate relatives, the terms standing for non-natural (or artificial) kinds. I will side-step this issue for the moment,[143] as my focus is on terms which are part of theoretical identifications and featuring in scientific theories. I will therefore accept that natural kinds are abstract objects, to maintain the strongest parallel between proper names and

[139] Armstrong [1997], p.106.
[140] Dupré [1995], p.20.
[141] Donnellan [1983], pp.90-91.
[142] Beebee and Sabbarton-Leary [2011], p.11. Similar criticism also in Soames [2002] and Schwartz [2002].
[143] Until chapter 7.

natural kind terms. If we make this assumption, Kripke's arguments for the rigidity of natural kind terms passes one hurdle. But there is another left.

2.8 Kripke and Kuhn's Challenge

2.8.1 Conceptual Change

The remaining hurdle is posed by conceptual change. Kripke does not discuss this explicitly in either *N&N* or "I&N." However, if we look closely at his emphasis on language constancy in thought experiments, we can extrapolate a stand on conceptual change, which has important consequences for the scope of his modal and semantic machinery.

When Kripke describes various thought experiments, he is always careful to insist that what is at issue is not how people in counterfactual situations would use certain terms, but rather what *we* should say, using *our* language, about those situations. For example, about the Hesperus-Phosphorus identity, Kripke asks what the conclusion would be if we imagined another world where people

> used the names 'Phosphorus' for Venus in the morning and 'Hesperus' for Mars in the evening…[W]ould it be a situation in which Hesperus was not Phosphorus? Of course, it is a situation in which people would have been able to *say*, truly, 'Hesperus is not Phosphorus'; but we are supposed to describe things in our language, not in theirs.[144]

We describe the situation in our language, not the language that the people in that situation would have used. Hence we must use the terms 'Hesperus' and 'Phosphorus' with the same reference as in the actual world. The fact that people in that situation might or might not have used these names for different planets is not relevant.[145]

Similarly, Kripke writes:

[144] Kripke [1971], p.155.
[145] Kripke [1981], p.109 note 51: "Recall that we describe the situation in our language, not the language that the people in that situation would have used."

> Suppose that all the areas which actually contain gold now, contained pyrites instead…Would we say, of this counterfactual situation, that in that situation gold would not even have been an element (because pyrite is not an element)? It seems to me that we would not…(Once again, whether people counter-factually would have *called* it 'gold' is irrelevant. *We* do not describe it as gold.) [146]

There is an implicit acknowledgement in these passages that terms can differ in different possible worlds. One illustration of the issue is given by Christopher Hughes, who quotes an old riddle. It imagines a riddler who asks his victim "If 'leg' meant 'tail-or-leg', how many legs would a horse have?" If this victim answers: "Five", the riddler responds "No, four: calling a tail a leg doesn't make it so." According to Hughes, this shows why Kripke emphasises that "we use our actual words, with their actual meanings, to describe counterfactual possible situations."[147]

However, I think the implications for natural kind terms go further than these simplistic examples indicate. It matters enormously that Kripke's theory is put forward as applicable only *given our current language*. It is not merely that it makes no difference to his theory if other proper names had been given to refer to Venus, or if the names that were actually given instead had been used to refer to Mars. The fact that Kripke's machinery only works on the condition that we restrict ourselves to our own language means that *the theory cannot be used as a reply to Kuhn's Challenge* (nor is there any sign that Kripke thought it could). If a change to a natural kind term is introduced, not by a switch of a word, but by a modification to its meaning prompted by a theory-change, Kripke's condition would be violated: we would no longer be using "our language", but a future language. The influence of such a change is not only relevant for scientists, because a change of scientific theory influences our vernacular terms and how we think about objects.[148]

There is one place in *N&N* that suggests that this is what Kripke has in mind:

[146] Kripke [1981], p.124.
[147] Hughes [2004], p.185.
[148] I will return to this relationship in Chapter 4.

> *[P]resent* scientific theory is such that it is part of the nature of gold as we have it to be an element with atomic number 79. It will *therefore* be necessary and not contingent that gold be an element with atomic number 79.[149]

The implication is that the necessity of the identity between water and H_2O presupposes the relevant scientific theory. If the theory changes, so too might the necessity. Natural kinds turn out to have a timeline. Kripke's theory so far does not give a response to Kuhn's Challenge.

2.8.2 Historical Chains

Maybe we can extrapolate a response to Kuhn's Challenge from Kripke's discussion of how proper names retain their reference, assuming that he would apply the lesson to natural kind terms. As we have seen, the view is that after an initial baptism, by ostension or description, the name is passed on to the next speaker with the intention to use it with the same reference. This suggests a simplistic historical chain model, where an unbroken chain is guaranteed by the speaker's intention only.

This account is untenable in its simplistic form, however, as pointed out by Gareth Evans. Evans takes the example of 'Madagascar', where the reference shifted from the African continent to the large island to the east, due to Marco Polo's misunderstanding.[150] Here the intention to use the word in the same way as the previous speaker (on the assumption that this was indeed Marco Polo's intention) does not guarantee an unbroken chain.[151] Kripke discusses Evans's point in the Addendum to *N&N*, where he adds a component to his theory, reducing the role of the historical chain:

> [A] present intention to refer to a given entity…overrides the original intention to preserve reference in the historical chain of transmission…[T]he phenomenon is perhaps

[149] Kripke [1981], p.125, italics added.
[150] Evans [1973], pp.195-196.
[151] Applied to natural kind terms in science, the simple chain model seems even less promising. Andreas [2017], §3.1.1., makes a Kuhnian point when he writes: "Kripke's story is particularly counterintuitive in view of the ahistorical manner of teaching in the natural sciences, wherein the original, historical introduction of a theoretical term plays a minor role in comparison to up-to-date textbook and journal explanations. Such explanations are clearly of the descriptivist type."

> roughly explicable in terms of the predominantly social character of the use of proper names…[W]e use names to communicate with other speakers in a common language. This character dictates ordinarily that a speaker intend to use a name the same way as it was transmitted to him; but in the 'Madagascar' case this social character dictates that the present intention to refer to an island overrides the distant link to native usage.[152]

Kripke is making a semantic point. The historical chain is the normal practice for how the reference of proper names is passed on, he claims, but this *can be overridden by the linguistic society*.

If Kripke intends that the same model is applied to natural kind terms, the conclusion would be that later usage by the linguistic community can override the historical function of a term. For example, even if 'whale' picked out a fish in the past, it needs not do so once we move from species classifications relying on similarity to those relying on genetic ancestry. In light of this, the appeal to a Kripkean historical chain is clearly not an argument that can be used against Kuhn's Challenge. Kuhn's point is not that there is *never* any continuity. He claims that when theories replace each other, there often occur meaning changes of key terms. Nothing in the extrapolated historical chain response counters that conclusion.

2.9 Conclusions

As I mentioned in the introduction, the Kripke-Putnam semantics has often been used as potential responses to Kuhn's challenge. To begin discussing whether this is correct, I introduced Kripke's machinery from *N&N*, starting with the application to proper names and then extended to natural kind terms.

A problem appeared immediately: if the extension of natural kind terms such as 'gold' and 'tiger' is the set of individual gold pieces and tigers, the crucial Kripkean tool of rigidity cannot be applied. But it is possible to construct a picture where the similarity is close if we assume that natural kinds are single, abstract objects. As name-equivalents, natural kind terms are rigid, and

[152] Kripke [1981], p.163.

identity statements between two natural kind terms are necessarily true, for the same reason as those containing proper names. With Donnellan, I will assume that seeing natural kinds as abstract objects gives the Kripke-Putnam semantics "the best run for its money".[153]

I have further argued that there is little reason to think that Kripke himself intended his semantics to be an answer to Kuhn's Challenge, insofar as he repeatedly insisted that the semantics assumes our current language. Kripke's position can be compared to scientists as described by Kuhn, who learn their trade with the help of increasingly complicated textbook exercises. One effect of this method is a lack of awareness of the historical continuity; scientists are trained according to the best available knowledge within the current paradigm, which facilitates rapid progress made in their disciplines. But as a side-effect, scientists working within an earlier paradigm appear odd, bordering on the incomprehensible, to their latter-day colleagues – how could these famous names believe in such strange things and make such blatant errors?[154]

Kripke is one of those scientists, working with thought experiments as his tools. For Kripke's "our current language" (where horses have no more than four legs) we can substitute "our current paradigm" (where water is H_2O) – and if we do this, his machinery works.

Kripke's model of a historical chain from baptism along a chain of further use does not give a response to Kuhn's Challenge either. This is because he allows that later usage by the linguistic community can override the historical function of a term.

If I am right about this, Kripke does not aim to provide an extra-theoretical framework to meet Kuhn's Challenge, nor does his account offer one by itself. Nevertheless, there are parts of Kripke's machinery that can be used to construct a potential response to Kuhn's Challenge. Putnam picks up this idea and I will therefore discuss his account in the next chapter.

[153] Donnellan [1983], p.90.
[154] A benign and plausible version of incommensurability, I suggest. Kuhn's example is Aristotelian Physics.

3 Putnam and Kuhn's Challenge

3.1 Introduction

In the previous chapter, I talked about the Kripke-Putnam semantics as a machinery that has been used to respond to Kuhn's Challenge. Kripke's emphasis is on proper names, while Putnam further develops the arguments in the area of natural kind terms. In this chapter, I will present some of those arguments.

There are many similarities regarding the main components of Kripke's and Putnam's machineries. They include their views of rigidity, necessity and chains of reference, and their criticism of the theory that Kripke calls "descriptionism". Putnam's criticism of descriptionism is in the same line as Kripke's, although he stresses that there are major social and environmental components to the meaning of natural kind terms. A similarity that I will focus on is their use of thought experiments. Putnam's "The Meaning of 'Meaning'" ("MoM")[155] contains a series of them, including the famous Twin Earth experiment. But I will also discuss more generally what type of thought experiments these are, what constraints one should put on their use, and which type of conclusions one can legitimately draw from each type.

There are also differences between the approaches of Kripke and Putnam. One difference is of course that Putnam later modified his views in many respects. In particular, this is true for his opinion on realism, which I discuss in Chapter 5. The view represented in *this* chapter builds primarily on "MoM" and its precursor "Is Semantics Possible?"[156]

A second difference is that for Putnam, meaning is defined as a combination of four elements: a syntactic marker, a semantic marker, a stereotype and the reference or extension. This is how it works for 'water':[157]

[155] Reprinted in Putnam [1975] (chapter 12), pp.215-271.
[156] Reprinted in Putnam [1975] (chapter 8), pp.139-152.
[157] Putnam [1975], p.269.

SYNTACTIC MARKERS	SEMANTIC MARKERS	STEREOTYPE	EXTENSION
mass noun; concrete	natural kind; liquid	colourless; transparent; tasteless; thirst-quenching; etc.	H_2O (give or take impurities)

I will discuss the latter two components of meaning in the next section. Later I will add to the picture a related Putnam term, 'the linguistic division of labour'.

A third difference, central to the present thesis, is that Putnam is firmly committed to the use of his semantics as an *extra-theoretical* tool, effective against Kuhn's Challenge. I argued in the previous chapter that it is at least doubtful that this is Kripke's aim. In "MoM", in contrast, it plays a major role. In this chapter I will outline the two types of responses to Kuhn's Challenge mentioned in "MoM", including Putnam's arguments for how conceptual continuity is possible despite meaning changes. I will identify two assumptions on which Putnam's arguments rely: the validity of thought experiments over time and possible worlds, and extra-theoretical essentialism. I will come back to both assumptions many times later on.

3.2 Reference Across Paradigms

As mentioned above, one element of meaning for Putnam is the *stereotype* associated with the term, already introduced in the last chapter. The stereotype is constituted by descriptions we typically associate with a term, and therefore serves to define what we need to know to be able to use that term correctly.

The notion of stereotypes is an improvement on Kripke's theory. It is consistent with Kripke's idea of reference via contingent properties but goes beyond, to complement this with a theory of linguistic competence. The separation of linguistic competence from what ultimately determines extension does the same job as the corresponding distinction in *N&N*: it facilitates the necessary *a posteriori*, where a necessarily true theoretical identification, defining a natural kind, is not

known to the speaker, who nevertheless successfully refers to this kind deploying the stereotype. Knowledge of the properties that determine the extension *can* be part of a stereotype, but it does not *have to be* for the stereotype to work.[158]

The other important component of meaning is the referent or extension. One consequence of this is that for Putnam, meaning includes reference by definition. But reference to what? Putnam says this about 'gold':

> [W]e maintain: 'gold' has not changed its *extension* (or not changed it significantly) in two thousand years. Our methods of *identifying* gold have grown incredibly sophisticated. But the extension of χρυσός in Archimedes' dialect of Greek is the same as the extension of *gold* in my dialect of English.[159]

The point is not that the total amount of the metal has stayed the same. There might not have been any major changes in how many samples of gold there are over time, but Putnam's argument should not depend on this. I will therefore pick up the idea used in the last chapter, that natural kind terms refer to abstract objects, and understand the extension of 'gold' as a natural kind, and thus an abstract object. Putnam's thesis then becomes:

> P1: The abstract object (the natural kind) that 'gold' refers to has not changed significantly since antiquity.

If the stereotype has stayed the same,[160] this also implies that:

> P2: The meaning of 'gold' has not changed significantly since antiquity.

It is certainly possible that Putnam is right about 'gold'. It is likely that there are terms where there is no major change in either meaning or reference over long times, and across theories, as

[158] Compare Kripke's main point against Frege: the (alleged) mistake to combine into one notion, the Fregean sense, the role of making it possible to understand what a term means with the role to determine its reference.
[159] Putnam [1975], p.235.
[160] Which in itself is open for doubt; see §4.3.

Robin Findlay Hendry proposes.[161] But even if the same were true of all natural kind terms (which is improbable), to be able to handle theory shifts *without* meaning changes is not enough for a framework for extra-theoretical continuity. Such a framework must guarantee continuity of reference regardless of meaning, making meaning redundant when discussing continuity of reference. Thus Hendry's proposal is not a reasonable interpretation of what Putnam proposes in "MoM". Putnam's ambition is to produce an extra-theoretical argument in favour of conceptual continuity and realism, responding to Kuhn's Challenge.

In "MoM", Putnam argues for his realism only indirectly, from the disadvantages of what he sees as the alternative: operationalism. "The alternative view is that 'gold' *means* whatever satisfies the *contemporary* operational definition of gold."[162] But Putnam states that operational criteria are inadequate "for the application of *any* such word."[163] Operationalism, Putnam suspects, is motivated by antirealism. Antirealism, Putnam says, would have to give up some indispensable concepts, and this would stop us from describing scientific development in terms of progress, because

> if we are to use the notions of truth and extension in an extra-theoretic way (i.e. to regard those notions as defined for statements couched in the languages of theories other than our own), then we should accept the realist perspective to which those notions belong.[164]

In contrast, the antirealist is left without a response to Kuhn's challenge. Putnam claims that

> the antirealist does not see our theory and Archimedes' theory as two approximately correct descriptions of some fixed realm of theory-independent entities, and…he does not think our theory is a *better* description of the *same* entities that Archimedes was describing.[165]

[161] Hendry [2010].
[162] Putnam [1975], p.235.
[163] Putnam [1975], p.238. Italics added.
[164] Putnam [1975], p.237.
[165] Putnam [1975], p.236.

I will therefore take Putnam's line in "MoM" to be that there is a general, extra-theoretical continuity of the reference of natural kind terms, backing up scientific realism, and effective against Kuhn's Challenge. But why should we believe that regardless of theory changes, natural kind terms have kept their reference (let alone their meaning) constant over time? There are two types of arguments in "MoM" that can be construed as an attempt to respond to Kuhn by building on the Kripke-Putnam semantics. I will discuss them below and introduce more of Putnam's additions and variations to Kripke's machinery as I go along.

3.2.1 Necessity-Based Continuity

The first of these arguments is found in the Twin Earth thought experiment,[166] where Putnam assumes the existence of a planet far away, identical to Earth in all respects with one exception: what they on this planet, Twin Earth, call "water" has a different microstructure than water on Earth. On Twin Earth, this microstructure is not H_2O, but another chemical, described by a formula abbreviated as XYZ. Twin Earth water is indistinguishable from Earth water for our senses, fills the Twin Earth seas, lakes and rivers, and quenches thirst exactly like Earth water. Then the difference in microstructure is discovered, following a visit to Twin Earth by astronauts from Earth. The question is, would we say that what is called "water" on Twin Earth is really water? Putnam says "no", it is *not* water, because a liquid is water if and only if it consists of H_2O. Instead, Putnam states, we would say something like: "On Twin Earth the word 'water' means XYZ."[167]

Putnam now envisages the situation on the two planets in 1750, and the existence at the time of the Earth inhabitant $Oscar_1$ and his exact physical and psychological duplicate on Twin Earth, $Oscar_2$. Their (narrow) psychological states in respect to what they call "water" are identical: they experience, know and believe exactly the same things about the respective liquid. The two Oscars would not know anything about the underlying microstructures, as these are yet to be

[166] In Putnam [1975), chapter 12 ("MoM"). This thought experiment has generated a very large amount of comments. A selection is included in Pessin & Goldberg [2015], which also includes an introduction by Putnam, written twenty years after the original article.
[167] Putnam [1975), p.223.

discovered. But the chemical difference between the two liquids called "water" in the two locations ($water_E$ and $water_{TE}$ respectively) would be the same in 1750 as it is today. Therefore, Putnam concludes, any theory that states that (narrow) psychological states determine extensions must be false. And if the extension is different, the meaning is by Putnam's definition different too. He has an argument to support that, but this requires the introduction of one more technical term.

Putnam's term that corresponds to Kripke's rigidity is 'indexicality'. According to Putnam, it is implicitly assumed that when we use a natural kind term like 'water', we refer to all tokens with the same *nature* ("essence") as the water we in fact have here on earth.

Putnam regards indexing as being in the same business as Kripke's rigidity. Putnam writes: "Kripke's doctrine that natural-kind words are rigid designators and our doctrine that they are indexical are but two ways of making the same point."[168] Whether indexicality and rigidity really are "two ways of making the same point" has been disputed,[169] but this does not matter for my purposes. Putnam clearly believes and intends them to be, and "heartily endorse(s)" the application of the notion of rigidity to natural kind terms. Therefore I will use the term 'rigidity' also in connection with Putnam's argument. The rigidity of 'water', Putnam means, is established by the Twin Earth thought experiment.

The rigidity of 'water' (and of 'H_2O') suggests one type of continuity argument; I will call this "Necessity-Based Continuity". It goes like this: "The scientific identification 'water is H_2O' is necessarily true because it is true in all possible situations, due to the rigidity of the terms involved. Consequently, it is true over time." As Putnam puts it: "Once we have discovered that water (in the actual world) is H_2O, *nothing counts as a possible world in which water isn't H_2O*."[170] And if there is no possible world in which water is not H_2O, then there is no possible time either.

[168] Putnam [1975], p.234.
[169] See LaPorte [2004], p.43, for a differing opinion.
[170] Putnam [1975], p.233.

An obvious counterargument to this is that the meaning of 'water' might have changed. Putnam handles this counterargument by introducing a modal version of the Twin Earth thought experiment. The version I have discussed so far is an epistemic thought experiment, which assumes that we find out something new (albeit extremely surprising) about our actual universe. In the modal version, Putnam describes planets in a possible non-actual universe. He supposes there are two possible worlds, W_1 and W_2. In both of them, Putnam is given an explanation of the meaning of 'water' by means of the phrase "this is water" stated by somebody pointing to the content of a glass. But in W_1, which is identical with our actual world, the content consists of H_2O, and in W_2 the content is XYZ.

Putnam suggests that "there are two theories one might have concerning the meaning of 'water'", either that it is "*world-relative"* so that it picks out XYZ in W2; or that "water is H_2O in all worlds" including W2.[171] Putnam opts for the second theory. Introducing another notion, the equivalence relation, he writes:

> When I say '*this* (liquid) is water,' the 'this' is, so to speak, a *de re* 'this' – i.e. the force of my explanation is that 'water' is whatever bears a certain equivalence relation…to the piece of liquid referred to as 'this' *in the actual world*.[172]

Deploying the equivalence relation of "same liquid" ("same$_L$"), his formalization shows how indexicality is meant to work:

> (For every world W) (For every x in W) (x is water ≡ bears same$_L$ to the entity referred to as 'this' in the actual world W_1).[173]

Again, Putnam stresses that this indexicality agrees with Kripke's way of making the point: "we may express Kripke's theory and mine by saying that the term 'water' is *rigid.*"[174] Since it is rigid, it must designate the same thing across possible worlds and times.

---

[171] Putnam [1975], p.231.
[172] Putnam [1975], p.231.
[173] Putnam [1975], p.231.
[174] Putnam [1975], p.231.

In Putnam's thinking, natural kinds such as water have observational properties that are contingent, and hidden properties that are necessary. Sorting relations such as $same_L$ are intended to use the necessary, hidden ones. Two objects are similar in a certain respect, and they belong to a natural kind in virtue of having some *particular* properties in common. These are what Kripke regularly, and Putnam sometimes, call their "essential" properties, which explain the observational properties. The essential properties connect individuals with a natural kind, and they also connect instances of the natural kind across time and possible worlds.

> A natural kind *term*…is a term that plays a special kind of role. If I describe something as a *lemon*…I indicate that it is likely to have certain characteristics…but I also indicate that the presence of those characteristics, if they are present, is likely to be accounted for by some 'essential nature…'[175, 176]

To evaluate Putnam's necessity-based continuity argument requires examining the implicit assumption of unchanged essences. I consider this topic in §3.6. Before that, I consider Putnam's second response to Kuhn's Challenge: the historical chain-based argument.

3.2.2 Historical Chain-Based Continuity and The Division of Linguistic Labour

The historical chain-based argument states that we can bypass all (other) meaning components by relying on an unbroken chain of reference following an initial baptism. I presented this response in §2.8.2. But we saw that a pure version of the historical chain argument does not work, due to Madagascar and similar cases, where an intention to use the term as the previous speaker produces the wrong result. To address this issue, Putnam incorporates the "social character" Kripke identified in natural kind terms in his response to the Madagascar case, and represents this with his notions of stereotypes and the *division of linguistic labour*.

We recall that Kripke in the main text of *N&N* sees the speakers' intention to use a term in the same way as earlier speakers as one prerequisite for continued reference. The notion of stereotypes makes it possible for Putnam to give an improved account of how the intention of

[175] Putnam [1975], p.140.
[176] Kuhn [1990] makes a similar point, stressing the need for re-baptisms, which he calls "redubbings".

speakers come into the picture, according to which the intention is to use a term as it used in the linguistic community. Within these communities, Putnam believes there is a division of linguistic labour.

Putnam says that in most or all societies, some terms have a gap between expert knowledge, on the one hand, and the stereotype competence needed to use terms to successfully refer, on the other. He even speculates that this division of labour is a "fundamental trait of our species."[177] Communities of experts evolve with the development of scientific theories and methods. In 1750, Putnam writes, there were no experts on 'water' because there were none, in a modern sense, on *water*: that is, there was nobody who knew what water really was, and could reliably tell whether a sample was water or not. The same was true of 'gold' in Archimedes' time.

The division of labour for natural kind terms in the linguistic community, Putnam suggests, builds on a non-linguistic division of labour. Many buy and wear gold rings, but they do not need to be able themselves to reliably tell whether the ring is really gold – they have good reasons to trust the expert from whom they bought it. This is reflected in the language.

How is the labour divided? On the part of the general public falls the duty to know the stereotype in order to be a competent user of the term in normal discourse. Stereotypes can be stronger, giving good criteria, or weaker. On the part of experts falls the duty to know the criteria for belonging to the extension. With the help of the division of labour we can share the expert knowledge, without being experts ourselves:

> [T]he way of recognizing possessed by these 'expert' speakers is also, through them, possessed by the collective linguistic body, even though it is not possessed by each individual member of the body, and in this way the most recherche fact about water may become part of the social meaning of the word while being unknown to almost all speakers who acquire the word.[178]

[177] Putnam [1975], p.229.
[178] Putnam [1975], p.228.

This means that we can be linguistically competent, proficient in using a term in conversations, even if we do not know the correct, exact criteria for the application of the term. Such conversations are not about knowing facts, Putnam says in his 1990 introduction to "The Twin Earth Chronicles":

> My suggestion was that knowing the meaning of the word 'gold' or of the word 'elm' is not a matter of *knowing that* at all, but a matter of *knowing how*; and what you have to *know how* is to play your part in an intricate system of social cooperation.[179]

The point is not just that the required level of knowledge can differ between experts and laypersons. It is also that the layperson's use of a term is dependent on the experts' existence. Putnam cannot tell the difference between elms and beech trees, he confesses. This is one example of where the (narrow) psychological state does not determine the extension – the two varieties of trees appear the same for Putnam. With the division of labour, and given the weak stereotype related to this tree, Putnam can get away with talking about elms despite his rudimentary knowledge about them. There is a dependence of ordinary language use on that of experts: "We could hardly use such words as 'elm'…if no one possessed a way of recognizing elm trees".[180]

The division of labour addresses the issue of extension-fixing.

> Whenever a term is subject to the division of linguistic labor, the 'average' speaker who acquires it does not acquire anything that fixes its extension. In particular, his individual psychological state *certainly* does not fix its extension; it is only the sociolinguistic state of the collective linguistic body to which the speaker belongs that fixes the extension.[181]

[179] Pessin & Goldberg [2015], p xvi.
[180] Putnam [1975], p.227.
[181] Putnam [1975], p.229.

Putnam's analysis of the natural kind term 'gold' illustrates the division of labour, both linguistically and non-linguistically.[182] We benefit from experts not only for investment purposes, but also for the correct usage of the term 'gold'.

> [N]ecessary and sufficient conditions for membership in the extension, ways of recognizing if something is in the extension ('criteria'), etc. – are all present in the linguistic community *considered as a collective body*; but that collective body divides the 'labor' of knowing and employing these various parts of the 'meaning' of 'gold'.[183]

Putnam's elaboration of linguistic competence solves one problem that the simple historical-chain model faced; to refer successfully is not a matter of using the term in the same way as another speaker, but using it according to how it is used in a linguistic community, according to its social meaning. The knowledge about correct usage is held within the group of experts. But all in the linguistic community share the collective knowledge thanks to the division of labour.

There is, however, an issue with Putnam's claim that we depend on experts for meaningful usage of natural kind terms: There are periods when Putnam says that there *are* no experts. In these periods it seems as if the chain is broken. Putnam hints of a possible solution to this, and I will in the next section both analyse the problem and present Putnam's solution. This solution is the key to Putnam's historical chain-based argument considered as a reply to Kuhn.

3.2.3 Division of Labour over Time

We found in the last section that the historical chain-based argument seems to be undermined by a dependence on non-existent experts. The problem is to reconcile these claims in "MoM":

1. To meaningfully use some terms, we are dependent on the existence of experts, through the division of labour. (Putnam clearly intends these terms to at least include terms for natural kind that have a microstructural essence, such as 'water'.)
2. Water was H_2O, and 'water' referred to H_2O, in the third century BC and in 1750.

[182] The example focusses on solid gold and of course ignores difference in language.
[183] Putnam [1975], p.228.

3. There were no experts on water, and therefore no experts on ‘water’, in the third century BC or in 1750.

Maybe Putnam does not hold that there must be a community of experts available *right there and then* for terms to be correctly used. He mentions another idea in passing: a division of labour over time.

> There were not (in our story, anyway) any ‘experts’ on water on Earth in 1750, nor any experts on ‘water’ on Twin Earth. (The example *can* be construed as involving division of labor *across time*, however. I shall not develop this method of treating the example here.)[184]

This idea might not be developed in “MoM”, but it is nevertheless implicit in Putnam’s case for extra-theoretical continuity: the primacy of our knowledge over those who lived earlier. Illustrating the point, Putnam imagines that Archimedes has discovered a sample of pyrite, but believes it to be gold, because he lacks the means to tell them apart. A kind time-traveller would have been able to put him right, Putnam feels, as the behaviour of gold and pyrite are similar but not identical.

> If, now, we had gone on to inform Archimedes that gold has such and such a molecular structure (except for X), and that X behaved differently because if had a different molecular structure, is there any doubt that he would have agreed with us that X isn’t gold? [185]

Would we have remaining doubts about Archimedes’ opinions, Putnam has a response to those too. Imagine that Archimedes insists that a particular sample of pyrite is gold, despite the time-traveller’s best efforts.

> Archimedes would have said that our hypothetical piece of metal X was gold, but would have been *wrong*. But *who’s to say* he would have been wrong?

[184] Putnam [1975], p.229.
[185] Putnam (1975), p.238.

> The obvious answer is: *we are* (using the best theory available today).[186]

In other words, what Archimedes was talking about relies on the contribution of present-day experts.

Division of labour over time is needed to prop up Putnam's continuity argument, but as a response to Kuhn's Challenge it leaves a lot to be desired. It seems to be begging the question by declaring a continuity, and a realism based on that, from our *current* position – which is exactly what Kuhn said scientists tend to do. Putnam is of course aware of that. But he argues that this is legitimate.

In §3.2.1 I introduced what Putnam calls "a certain equivalence relation", which in the case of water is called "same$_L$". As his intended conclusions are not restricted to water, I will now start using a general form of this relation: *same$_x$*. The same$_x$ relation is a "theoretical relation", Putnam says. Whether a sample is the same$_x$ as another one "may take an indeterminate amount of scientific investigation to determine."[187]

For Putnam, it is the same$_x$ relation that justifies stating that we are indeed talking about the same thing, gold, as Archimedes; we do so in virtue of the same$_x$ relation that binds samples of metal together over time. The experts that the linguistic community in Greece 300 BC benefitted from, then, are the chemists of today. Note, however, that the confidence in the superior diagnostics available now has the prerequisite that Archimedes, when uttering "water", intentionally referred to samples that not only had a hidden structure, but whose *essential property* was this hidden structure. Further, the claim that Archimedes might agree that he was wrong about a sample of pyrite presupposes that Archimedes himself *believed* that the samples had a hidden essential structure. This is what Putnam argues for in "MoM". It looks like attributing improbable beliefs to people in antiquity for the purpose of saving a theory. In Chapter 7, I will argue that the starting point of Putnam's argument, namely the trust he puts in later experts with a division of labour over time has merits on its own (and that it is not a million miles away from Kuhn's own

[186] Putnam [1975], p.236.
[187] Putnam [1975], p.225.

considered position), and (in Chapter 6) that the historical chain argument can be saved. Fortunately, we do not need implausible belief attributions for this.

I have now introduced Putnam's arguments for conceptual continuity. In the rest of this chapter I will present two main assumptions that these arguments rely on, his auxiliary hypotheses, so to speak. These are: the validity of conclusions drawn from thought experiments, and the existence of essences over time and across possible worlds. I will claim that though the former is defensible, the latter is problematic.

3.3 Putnam's First Assumption: The Validity of Thought Experiments

The first assumption, which Putnam shares with Kripke, is a dependence on the validity of conclusions drawn from philosophical thought experiments. I will discuss several issues with this assumption. Although the objections have merit, thought experiments remain very useful tools, which justifies Putnam's reliance on them.

In one type of philosophical thought experiment, Kripke and Putnam vary our epistemological situation by asking us to imagine that we find out something we did not know before (the *epistemic* version). In another type, they put us in a counterfactual situation, where the situation in some specified way differs from the real world (the *modal* version). Typically, these situations are very similar to our actual situation, but with some important exception.

Thought experiments, in this sense, are exercises in conceptual analysis. For Kripke and Putnam, linguistic intuitions carry great weight. What would be announced? What would we say? Putnam writes: "[I]f the color of lemons changed – say, as the result of some gases getting into the earth's atmosphere and reacting with pigment in the peel of lemons – we would not say that lemons had ceased to exist".[188]

In the last chapter, I related the experiment where Kripke imagines we have been suffering from an illusion regarding the colour of gold – it is in fact blue. What would we say when we find out?

[188] Putnam [1975], p.142.

> [W]hat would be announced would be that though it appeared that gold was yellow, in fact gold has turned out not to be yellow, but blue.[189]

Putnam describes the choice between denying either that intensions are determined by psychological states or that intensions determine extensions, and he goes for the first choice – based on linguistic intuitions:

> Consider 'elm' and 'beech', for example. If these are 'switched' on Twin Earth, then surely we would *not* say that 'elm' has the same meaning on Earth and Twin Earth, even if my *Doppelgänger's* stereotype of a beech (or an 'elm', as he calls it) is identical with my stereotype of an elm. Rather, we would say that 'elm' in my *Doppelgänger's* idiolect means *beech*.[190]

With the help of our linguistic intuitions in such situations, that is, with the help of conceptual analysis, we can identify which terms that are rigid, those that refer to the same person or object in all possible worlds.

One question, though, is how stable and reliable such linguistic intuitions are. Daniel Dennett questions this when he talks about certain thought experiments as "intuition pumps": "Intuition pumps are cunningly designed to focus the reader's attention on 'the important' features, and to deflect the reader from bogging down in hard-to-follow details."[191] This can be a good thing, Dennett stresses, but "intuition pumps are often abused, though seldom deliberately."[192]

Dennett's scepticism has particular force when the use of a term is stretched outside its normal domain of use. The discussion of personal identity is well-known for its outlandish examples, and the case against the reliability of linguistic intuitions correspondingly strong. With changes in the description, intuitions can change direction.[193] But in the cases of Nixon and Aristotle that

[189] Kripke [1981], p.118.
[190] Putnam [1975], pp.245-246.
[191] Dennett [2015], p.13.
[192] Dennett [2015], p 13.
[193] This is Dennett's argument against Searle's Chinese Room experiment (Searle [1980]); Dennett wants to show that he can generate different intuitions than the ones Searle generated. He writes in Dennett [1980], p.429: "In

Kripke discusses, it is difficult to believe that the scenarios discussed would be impossible in any sense, nor do they feel strange or unfamiliar. We can easily relate to scenarios where Nixon lost an election he actually won, and a scenario where we find out that Aristotle never taught Alexander. These are the types of counterfactual reasoning that we habitually engage in. Similarly, we can without too much of a stretch imagine gold being blue.

For Kripke, thought experiments have a high epistemic value because he sees a very close connection between the conditions according to which we correctly call something "X" and the conditions according to which we are justified to say that something is an X.

> Of course, some philosophers think that something's having intuitive content is very inconclusive evidence in favor of it. I think it is very heavy evidence in favor of anything, myself. I really don't know, in a way, what more conclusive evidence one can have about anything, ultimately speaking.[194]

There are nevertheless, in Kripke's opinion, constraints to what we are allowed to intuit from thought experiments. Their conclusions are strong indications, but they are not infallible. Let us turn to the issue of what these constraints are, and what implications follow.

3.4 What We Can and Cannot Think

Neither in epistemic nor in modal thought experiments can we postulate something that is impossible, Kripke and Putnam insist. If something cannot exist, we cannot find it in an epistemic thought experiment, nor postulate it in a modal one. The states-of-affairs in both types of scenario must in principle be such that they can occur – they must be possible. The forbidden impossibility is not just logical, but also metaphysical. We cannot postulate the existence of a square circle. But nor can we postulate a situation where water is not H_2O. Necessary (essential) properties are the properties that an object cannot be missing, because it would then no longer be that object. It is not just that we are told not to do it; we cannot, *however hard we try*, imagine

this instance I think Searle relies almost entirely on ill-gotton [*sic*] gains: favorable intuitions generated by misleadingly presented thought experiments."

[194] Kripke [1981], p.42

something impossible, like imagining an object without its essential properties.[195] We can *believe* that we do, but we fool ourselves. What we in fact imagine is an *epistemological equivalent*. Take a situation where a liquid is similar to water in all respects we can think of, except for its microstructure. In that situation, we can believe that we imagine non-H_2O water. But we cannot *truly* imagine that water is not H_2O, because if something is not H_2O, it is not water; H_2O is the essence of water. Similarly, we can imagine gold being blue, but not gold having another atomic number than 79, because it would be another substance. Would it for example have atomic number 47 instead, it would be silver. I will come back to this topic in §4.3.

Leibniz's Law stipulates that if A and B are identical, they must have all their properties in common. A general problem with thought experiments is that neither when we talk about identity over time, nor when we talk about identity across possible worlds, do the identity relations used obey Leibniz's Law. Both 'Richard Nixon the future US President went to Harvard' and 'Richard Nixon could have lost the 1968 election' assume that the young Richard is identical with the middle-aged man, and that the election-winner is identical with the election-loser, despite obvious differences in properties.

As we saw in the last chapter, Kripke regards worries about identity across possible worlds, relevant for modal thought experiments, as a pseudo-problem.

> [T]he man who might have lost the election or did lose the election in this possible world is Nixon, because that's part of the description of the world. 'Possible worlds' are *stipulated*, not *discovered* by powerful telescopes. There is no reason why we cannot *stipulate* that, in talking about what would have happened to Nixon in a certain counterfactual situation, we are talking about what would have happened to *him*.[196]

For Kripke, trans-world identification is unproblematic, because Nixon's persistence over time and possible worlds is a matter of stipulation. But this is not immediately convincing, If we insist that we are speaking about *him*, there must in the actual world be a *him* to speak of, that is, there

[195] I suggest this needs to say "at pain of inconsistency".
[196] Kripke [1981], p.44.

must have been a Nixon in the actual world. Furthermore, the existence of Nixon in other possible worlds must be possible. Would we believe that persons across time is a fiction, as Derek Parfit does,[197] or that the actors in possible worlds are our counterparts rather than us, as David Lewis does,[198] the conclusion we believe we can establish by thought experiments would be unjustified. Rigidity is a semantic property of certain terms. We cannot directly draw metaphysical conclusions from semantic facts.

Kripke mentions one supporting argument: as we saw, he regards the linguistic intuition as "very heavy evidence in favor of anything".[199] If my reading of Kripke is right, this is a stronger argument than it might look. I argued in the previous chapter that Kripke does not intend his semantics to have extra-theoretical (cross-paradigm) implications; his conclusions are only applicable *given our language*. With this proviso, the linguistic intuitions represent how we see the world and what we know about it,[200] and for *this* reason they are "very heavy evidence". On that basis, I suggest that *it is not Kripke* who stipulates his use of the proper name 'Nixon', thereby making himself immune from requirements of identity across time and possible worlds. That move would be arbitrary and weak, as such an identity is controversial. Instead, Kripke captures the way we talk and think about persons over time and in counterfactual situations; we assume that there is a person such that he can survive through time and meaningfully feature in counterfactual situations. Kripke is not *making* a stipulation, he is *discovering* one. Kripke's claim is that we all stipulate that we are talking about Nixon across time and possible worlds when we use the name 'Richard M. Nixon', or variations thereof, intending to refer to the person who (in the actual world) was the 37th president of the United States, in historical and counterfactual situations. This claim is primarily *semantic*; it is about how we use proper names in epistemic and modal thought contexts. When we use them, we implicitly commit to the

[197] Parfit [1971], p.25: "If I say, 'It will not be me, but one of my future selves,' I do not imply that I will be that future self. He is one of my later selves, and I am one of his earlier selves. There is no underlying person who we both are."
[198] Lewis [1971], p.205: "To say that something here in our actual world is such that it might have done so-and-so is not to say that there is a possible world in which that thing *itself* does so-and-so, but that there is a world in which a *counterpart* of that thing does so-and-so."
[199] Kripke [1981], p.42.
[200] With Kuhn's term: our "world view".

legitimacy of identity across time and possible worlds, or rather, we make use of the customary stipulation that this is the case.

Putnam is, however, in a different situation than Kripke, because his theory has extra-theoretical ambitions. When he discusses the modal version of the Twin Earth experiment, Putnam says: "We shall assume further that in at least some cases it is possible to speak of the same individual as existing in more than one possible world."[201] But he immediately adds a footnote: "This assumption is not actually needed in what follows. What *is* needed is that the same *natural kind* can exist in more than one possible world."[202] For Putnam, who wants to give an answer to Kuhn's Challenge, a justification for linguistic intuitions becomes an issue, and he relies on his second major assumption to address this, his essentialism.

Essentialism plays a role in another restriction on what we can do in thought experiments, indicated by Kripke with his unicorn example. Although there are no unicorns, could there have been some? Or is it necessarily true that there are none? Kripke rejects both these positions.

> Perhaps according to me the truth should not be put in terms of saying that it is necessary that there should be no unicorns, but just that we can't say under what circumstances there would have been unicorns.[203]

Kripke expands on this example in his Addenda to *N&N*, making it clear that what he thinks is missing is a sufficient level of detail about unicorns' *essence*, about what makes a beast into a unicorn.[204] There can be no natural kinds without essences, assigned to the natural kind object at its baptism. I will discuss the role of essentialism for Putnam's argument in §3.6.

3.5 Consistency and Completeness

There are two more question marks about the basis for conclusions from thought experiments. The set of postulates in a thought experiment, it is often said, should be *complete* and *consistent*.

[201] Putnam [1975], p.230.
[202] Putnam [1975], p.230, footnote.
[203] Kripke [1981], p.24.
[204] Kripke thinks about essences as microstructures, but I leave this discussion until the next chapter.

I will first discuss a worry about the completeness requirement for modal thought experiments, and then a potential issue related to the consistency requirement for epistemic thought experiments. In the end, I will suggest that the consistency issue can be eliminated, though we can and must live with the completeness worry.

Teresa Robertson and Philip Atkins describe the completeness requirement in the following way. "Philosophers typically regard possible worlds as giving a complete description of a possible state of the universe."[205] This is an impractically strong requirement, and Scott Soames puts it even more strongly: "Metaphysically possible world-states are maximally complete ways the real concrete universe could have been – maximally complete properties that the universe could have instantiated".[206]

Kripke notices the problem. In the First Lecture of *N&N*, he says, "[I]n theory, everything needs to be decided to make a total description of the world."[207] But he is not too concerned, because

> in practice we cannot describe a complete counterfactual course of events and have no need to do so. A practical description of the extent to which the 'counterfactual situation' differs in the relevant way from the actual facts is sufficient…[208]

I believe that *there is* a response to the completeness problem in *N&N*, although it is easy to miss, and that this response has implications for how we should look at thought experiments. This answer is that the description of a possible world is compared, not with the real universe out there, but with *a chosen set of relevant descriptions* of the actual world. If I am talking about the effect on US foreign policy in a possible world where Richard Nixon lost the 1968 election, I do not worry about the effect this had on Nixon's personal affairs, because I do not include anything about Nixon's family in my descriptions of the *actual* situation. Possible worlds *do* need to be complete, but only relative to the chosen descriptions with which they are compared. Kripke writes: "The 'actual world'…should not be confused with the enormous scattered object that

[205] Robertson and Atkins [2018], §2.
[206] Soames [2011], p.80.
[207] Kripke [1981], p.44.
[208] Kripke [1981], p.18.

surrounds us."[209] The 'actual world' is here the rather misleading term for a set of descriptions that cover a sub-set of what we hold to be features of that "scattered object" – the real world we live in. So interpreted, the completeness requirement looks reasonable, but it leaves an in-built question for modal thought experiments to address. Kripke writes: "A practical description of the extent to which the 'counterfactual situation' differs *in the relevant way* from the actual facts".[210] It must be decided which descriptions of the actual world that are relevant to include.

Epistemic thought experiments appear to have issues with the consistency requirement. It is implicit in an epistemic thought experiment that everything is unchanged compared with the actual world, except the imagined discoveries described in the experiment. The idea is that what is described might exist, although we do not know about it. We could one day find out. It does not matter how unlikely it is that we do, as long as it is possible. The main version of Putnam's Twin Earth is an epistemic thought experiment and there are indeed some question marks related to the its consistency. As I said, in an epistemic experiment, we expect no revision of our standards, just an extension to a situation that in principle could occur. But on Twin Earth, we do not seem to be routinely applying the usual meaning of a term to a normal science situation.

Elke Brendel gives some details about what is missing:

> [B]y varying one particular factor of our world in his imagination (water is no longer H_20 but XYZ), Putnam fails to pay attention to the drastic effects this variation has for twin earth and its inhabitants. He merely, and illegitimately, stipulates that everything else remains the same. But of course, if the liquid on twin earth is not H_20 our "twin earth Doppelgängers" cannot be molecularly identical to us. About 70 % of a human being consists of H_20 molecules. If we exchange an important chemical substance with something else, the so-called twin earth will be completely different from the world we live in and – contrary to what Putnam will have us believe – we will have not the slightest

[209] Kripke [1981], pp.19-20.
[210] Kripke [1981], p.18 [italics added].

idea of what this strange world and the psychological states of its inhabitants (if they have any) will be like.[211]

Brendel talks about underdescription,[212] which is a worry about completeness, but the problem might run deeper than that. Kuhn remarks that the real reaction to a Twin Earth discovery would be to question fundamental chemistry, as no two substances *can* have identical observational properties. He insists that rather than "On Twin Earth, the word 'water' means XYZ," Putnam's astronauts would report: "Back to the drawing board! Something is badly wrong with chemical theory."[213, 214]

If an epistemic thought experiment, as a natural reading suggests, assumes a *ceteris paribus* clause, the Twin Earth thought experiment is not consistent. Everything else *cannot* be the same if we postulate the existence of XYZ with exactly the same observational properties as H_2O.[215, 216] But the conclusion, I suggest, should not be that the Twin Earth experiment is *inconsistent*. A better and more charitable interpretation is that the experiment, as Brendel suggests, is *incomplete*. I would add that this incompleteness is defendable. It is the *ceteris paribus* idea that is mistaken.

Brendel points out that Putnam states that "everything else stays the same", implying a *ceteris paribus* clause, and that this is illegitimate. As she says, the implication of the XYZ story would go far beyond what Putnam describes. But I do not agree with Brendel that this incompleteness is fatal, or even avoidable – if we keep in mind the nature of the reference set used for comparison. The situation is the same as for modal thought experiments. Descriptions must in practice *always* be incomplete, and the selection we choose to include will reflect our purpose of enquiry.

---

[211] Brendel [2004], p.98.
[212] Kathleen Wilkes [1988] also points to the problem with underdescribed thought experiments, using examples from the debate about personal identity.
[213] Kuhn [2002], p.80.
[214] Dupré [1995], p.26, suggests that the fact that these experiences tend *not* to happen is why microstructure often is regarded as a promising candidate as an essential property.
[215] The problem turns up for in another form for modal experiments. If we postulate counterfactual chemical relationships, the experiment can be seen as incomplete, unless we supply a chemical theory.
[216] This is not a problem for the discussion of gold, where Putnam makes a point about the subtle differences between gold and pyrite.

As I said, we can describe the effect of a Nixon defeat in the 1968 election without saying a word about the effect on his private life. We compare the counterfactual situation with a description of a chosen subset of the actual world, not with the complete thing, whatever a 'complete' description of the world might mean. The selection, not just of the furniture of the counterfactual world, but also of the model actual world, is of critical importance. We choose according to our interests and purposes, which determine what is "the relevant way". This is also the case for epistemic thought experiments. We cannot describe either the current world or any alternative scenario in fully consistent detail. Nor can we know for certain the full implications of the variation we are introducing.

Having said that, completeness issues do certainly not make thought experiments useless. They can be powerful and informative even if entire theories are lacking. James Robert Brown and Yiftach Fehige write:

> We learn a great deal about the world and our theories when we wonder, for instance, what would have happened after the big bang if the law of gravity had been an inverse cube law instead of an inverse square. Would stars have failed to form? Reasoning about such a scenario is perfectly coherent and very instructive, even though it violates a law of nature.[217]

I conclude that none of these complications is fatal to the thought experiments Putnam, like Kripke, relies on. In the next section I turn to Putnam's second assumption, which is more problematic.

3.6 Putnam's Second Assumption: Extra-Theoretical Essentialism

Stathis Psillos formulates a requirement for continuity across conceptual change from a realist's perspective. The similarity between the earlier and the later concept, Psillos, says, must include substantial overlaps of *core properties* to count as the same concept. The requirement cannot be for *all* properties to overlap, because then we do not have any continuity, as Kuhn showed. But it

[217] Brown and Fehige [2019], §4.

cannot be just *any* property either, as Psillos wants to distinguish cases where there is continuity with a change in properties (such as electrons from Bohr until today)[218] from cases where a term is deemed not to refer (such as 'phlogiston'). The required overlap relates to *some* features only, the core properties.

Psillos's term 'core properties' is very similar to what I have understood by "essential properties", and his requirement for scientific continuity is in effect a requirement for unchanged essences. This is also what Putnam has in mind in his argument for extra-theoretical conceptual continuity that I discussed in §§3.2.1-3.2.3. In these sections, I introduced Putnam's key relation $same_x$ that he utilises in his continuity arguments to express what it takes for two samples to belong to the same natural kind. For $same_x$ to work across time and possible worlds, there must be objects that exist over time and possible worlds. For a natural kind, this is the continued existence of the abstract object in virtue of the essential properties that define it.

In this section, I will discuss a Putnam's version of extra-theoretical essentialism and point to a further assumption, the continuity of essence type. This will first be brought out by another look at the necessity-based argument for conceptual continuity, first introduced in §3.2.1, starting with a first, crude version:

> P0. It is by definition necessarily true that a sample consists of water if it has the same essence as a paradigm sample of water.
>
> P1. The paradigm sample consists of the chemical H_2O.
>
> P2. This sample consists of H_2O.
>
> Conclusion: It is necessarily true that this sample consists of water.

But there is an issue with this argument: as it stands, the conclusion does not follow from the premises. To be valid, it needs what Nathan Salmon calls "a non-trivial essentialist import": the assumption that water has a *chemical* essence. I will refer to this as an assumption about *essence*

[218] See §5.4.4.

type, the category that the essence belongs to. If we add this essentialist premise, we get a valid argument.[219]

> P3. To be a sample of water is to consist of the same *chemical substance* as other samples.

P3 is implicit in Putnam's argument, but needed for the conclusion. The question now is: How do we know P3 is true? For this we need a theoretical framework with a classification system. Without a category – in this case, chemical substance – no necessary properties (essences) can be discovered; this is what Salmon tells us. A few years after "MoM", Putnam recognises this point, when he writes: "'[O]f the same kind' makes no sense apart from a categorical system which says what properties do and what properties do not count as similarities."[220] For an essential property, the category must be of the right type, the type that contains potential essences.

Putnam's version of the necessity-based argument in "MoM" is not as crude as the one I have outlined above. As we have seen, he refers to a requirement for "a certain equivalence relation", the $same_{x,}$ to hold between two samples if they are both samples of water. But we now know that to do the job, the $same_x$ relation must include an essence-type assumption. We can discover that two samples have a $same_x$ relation – have the same essence – only if we have already established that the property relevant for this relation is of a certain type, as the samples might have very many properties in common. We need to know what type of sameness counts, that is, we need to know the essence type.

For my analysis of Putnam's argument, I will introduce a rough temporal indexing to indicate the meaning of terms at a given time. To look at the Twin Earth experiments with temporal indexing,

---

[219] This is Salmon's own version in Salmon [1982], p.166:

> P1: It is necessarily the case that: something is a sample of water if and only if it is a sample of dthat (the same substance that this is a sample of).
> P2: This [the liquid sample] has the chemical structure H_2O.
> P3: Being a sample of the same substance as something consists in having the same chemical structure.
> Therefore
> C: It is necessarily the case that: every sample of water has the chemical structure H_2O.

[220] Putnam [1981], p.53.

I will distinguish five meanings of 'water': one for today; one for 1750, where Putnam places the experiment; one for 1150; and one for 250 BC. The 1150 date is picked to be well before the scientific revolution. In addition, I will talk about a year well in the future, 2075.

Using these indices, I can clarify the essentialist second premise in my version of a Salmon-type argument, determining the time-dependent meaning of the natural kind term 'water':

> P0. It is by definition necessarily true that a sample consists of $water_{Today}$ if it has the essence as a paradigm sample of $water_{Today}$.
>
> P1. The paradigm sample consists of the chemical H_2O.
>
> P2. This sample consists of H_2O.
>
> P3. To be a sample of $water_{Today}$ is to consist of the same *chemical substance* as the paradigm sample.
>
> Conclusion: It is necessarily true that this sample consists of $water_{Today}$.

If we start by looking at the situation in 1750, Putnam says that '$water_E$ is H_2O' is necessarily true, albeit unknown to anyone, because it is also *a posteriori* and as yet undiscovered. Using our indices, we can answer that the '$water_{Today\text{-}E}$' meaning certainly was not in the heads of speakers in 1750 when they were talking about water.[221] But this is not what Putnam means; he wants to say that *no* sense of 'water', including '$water_{1750}$', was in their heads. As we saw, Putnam argues that this generalisation is legitimate, as there has been no meaning change affecting 'water' between 1750 and today.[222] If we accept that the stereotype is the same, and that the molecules making up the oceans etc. are the same, the unchanged meaning follows – but only if we also add the important qualification that the relevant $same_x$ relation assumes that water has a chemical as its essence.

[221] It might be objected that it was at least partly the same, as the stereotype was shared. But I dispute that too, in §4.3.

[222] Kripke says something very similar about the 1750 case. But the way I construed Kripke's argument makes it different from Putnam's in both logic and scope (admittedly based on limited texts to work on). Kripke's view, I said, was that there was no meaning change between the period when the (hypothesis of) and essence type was accepted and the period when the essence was found. However, I criticise this position in §7.6.

Also in 1750, Putnam says, the meaning of '$water_E$' and '$water_{TE}$' differed, due to their extension: "*The meaning was different because the stuff was different*."[223] And vice versa: the meaning of '$water_{1750\text{-}E}$' and '$water_{Today\text{-}E}$', we should conclude, are the same (as are '$water_{1750\text{-}TE}$' and '$water_{Today\text{-}TE}$'), because the (respective) stuff is the same. But this needs the same qualifier, the assumption that water has a chemical as its essence type.

Putnam's elaboration of the necessity-based argument for extra-theoretical continuity for natural kind terms relies on an essentialist assumption: that the natural kinds have essences that are constant over time, keeping the meaning constant as well. Using the indices, he is saying that $water_{Today} = water_{1750} = water_{1150} = water_{250BC,}$ or near enough, implying that '$water_{Today}$' = '$water_{1750}$', etc. This in turns requires an essence type assumption, the assumption that an object continues to have essences of a certain type. H_2O can only be the essence of water if water has a chemical as its essence type; this essence type will need to have been constant too to establish extra-theoretical continuity with the necessity-based argument.

To illustrate the point from another angle, I will now return to the historical chain-based argument for continuity, described in §§3.2.2.-3.2.3. This is the logic of the argument:

- Abstract object O_1 (a natural kind) has the name 'N' at time T_1 in a certain linguistic community, C_1. The stereotype of O_1 consists of two observational properties P_a and P_b.
- O_1 is not covered by any scientific theory at T_1, and there is therefore no expert knowledge of O_1 available.
- Speaker S_1 uses 'N' correctly at T_1 when they use it in the way it is used in C_1. This use is based on knowledge of P_a and P_b.
- At time T_2, when speaker S_2 lives in the linguistic society C_2, abstract object O_2 (a natural kind) has the name 'N' at time T_2 in a certain linguistic community, C_2. The stereotype of O_2 normally consists of two observational properties P_a and $P_{b,}$ but it has been discovered that some samples, *also* called 'N', have observational properties P_a and P_c.

[223] Introduction to Pessin & Goldberg [2015], p xvii.

- O_2 is covered by a scientific theory, which postulates that O_2 has the microstructural property MP as its essence, and there is therefore expert knowledge of O_2 available.
- Speaker S_2 uses 'N' correctly at T_2 when they use it in the way it is used in C_2. C_2 benefits from the existing expert knowledge, in a division of labour, and the identification of O_2 with E.

Now the question is whether 'N' at T_1 and 'N' at T_2 refer to the same object, so that O_1 and O_2 are identical. This is what the historical chain argument claims. But how do we know that the historical chain of reference to the object (O_1/O_2) is unbroken, and that we do not have another Madagascar case where the new social use of the term has changed the reference?[224] For natural kind terms, that is a recurrence of Kuhn's Challenge. To respond to that within Putnam's theory, we again need to rely on his idea that what guarantees continuity is a $same_x$ relation between O_1 and O_2. Because *$same_x$* relies on unchanged essences (within an essence type), we can say that object O_1 and O_2 are identical if and only if they have the same essential properties (both O_1 and O_2 have MP as their essence) and that therefore $O_1 = O_2$. We need earlier linguistic communities to use a $same_x$ relation with the same value – the same microstructural essence type – as later communities. It does not matter that there is variation in observational properties, as those are contingent. Nor does it matter if the essence was known to C_1. What matters for the identity are the essential properties. However, relying on essential properties presupposes an extra-theoretical agreement about their essence types to work, because essences can neither be discovered, nor wait to be discovered, without their essence type. It is not enough if we know that both O_1 and O_2 consist of MP. We also need to assume that O_1 had a microstructural essence type at T_1 so that MP could have been its essence.

Back on Twin Earth, the $same_x$ relation is operative when we use 'water' to refer to water across time and possible worlds. The quote below shows how it is used to establish the historical-chain argument. Putnam talks about repeat ostensive definition, rather than chains, but these keep pointing to the same thing:

[224] See §2.8.2.

> But, it might be objected, why should we accept that the term 'water' has the same extension in 1750 and in 1950 (on both Earths)? The logic of natural-kind terms like 'water' is a complicated matter, but the following is a sketch of an answer. Suppose that I point to a glass of water and say 'this liquid is called water'…My 'ostensive definition' of water has the following empirical presupposition: that the body of liquid I am pointing to bears a certain sameness relation…to most of the stuff I and other speakers in my linguistic community have on other occasions called 'water'.[225]

The "certain sameness relation", the $same_x$ relation, can stay the same only with the essentialist assumption. Also for historical-chain continuity we depend on natural kinds with constant essences and essence types.

Salmon therefore points to a weakness in Putnam's "MoM" argument, the implicit dependence on an essence type, indicating a requirement for unchanged essence types across paradigm shifts to make conceptual continuity possible. However, Putnam has a reply to Salmon: namely, that this requirement is trivial, because a "hidden" microstructural essence is the obvious choice, the only real alternative in cases like water. I will discuss this reply in the next chapter, where I will argue that Putnam is mistaken; the choice of essence type is often far from trivial.

3.7 Conclusions

The arguments for conceptual, extra-theoretical continuity that Putnam puts forward in "MoM" rely on two main assumptions: one is the validity of thought experiments and the other essences that exist over time and across possible worlds. He also needs the notion of a division of labour over time, which requires further support to avoid the impression of missing the point.

I discussed possible objections to philosophical thought experiments, but concluded that these can be answered and that thought experiments remain a useful tool. But they are always open to questions about the reliability and the source of our intuitions. Are we stretching the use too much? Are there counter-intuitions? Even more fundamentally, there are always issues of

[225] Putnam [1975], pp.224-225.

selection. What we include in the description of the universe under investigation is never complete, in the sense that it never completely maps facts about the physical universe. A description requires a selection, and the selection of what is relevant is not objectively given; it is a model that reflects *our interests*, the issues we want to investigate. As Kripke says: "Of course when we specify a counterfactual situation, we do not describe the whole possible world, but only the portion that interests us."[226] The same must be the case for the description of the actual world. The selection always runs a risk of being criticised for having omitted relevant descriptions, and it needs to be defended against accusations of underdescription and inconsistency. There is a similarity between this point and modelling done in science.

As for the second assumption, essences are properties that are the necessary and sufficient conditions for an object being the object it is. They define the abstract objects that are natural kinds. This is, according to my reading, a natural interpretation of Kripke's view. But Putnam needs a stronger version of essentialism, to support his extra-theoretical claims and his response to Kuhn's Challenge.

Following Salmon, we found that essences cannot be discovered without a choice of category from which an essence is to be found: an essence type. Putnam's $same_x$ relation, crucially important for his arguments for conceptual continuity, therefore requires not only the essence but also the essence *type* to be constant over time and across possible worlds, but Putnam regards this as unproblematic. In the next chapter I question this. I will discuss different essence types, the reason Putnam thinks he knows what Archimedes believes, and the relationship between vernacular and scientific terms.

[226] Kripke [1981], p.49, note 16.

4 Natural Kinds and Their Essences

4.1 Introduction

Kuhn's Challenge says that a change between two theories can only be progress if the two theories address the same subject matter. It is not straight-forward to specify what the condition "address the same subject matter" might mean without also putting forward a solution. But for natural kind terms we have already found that *some* proposed solutions are not promising. An example of one such solution is to require that natural kind terms of subsequent theories refer to the same individual entities, as they often do not – individuals can be created or perish over time. In Chapter 2, I therefore agreed with Donnellan that natural kinds are better seen as abstract objects. But if natural kinds are abstract objects, we cannot demand that the natural kind terms of subsequent theories refer to (abstract) objects with exactly the same properties, as a strict identity relation would require, for the reason Kuhn gives: definitions of natural kinds often change with new theories. This leaves us with the question of what the criteria for conceptual continuity are.

One type of answer to Kuhn's Challenge is built on the Kripke-Putnam semantics and in earlier chapters I outlined two main arguments, the necessity-based and the historical chain-based arguments. I found that both these responses rely on essentialist assumptions: that there are especially important properties, essences, that can persist over time and possible worlds, and that unchanged essences provide conceptual continuity. Furthermore, we found that this in turn implies that the category where a particular essence can be found, what I called its "essence type", also is unchanged. But Putnam argues in "MoM" that this is not a serious problem, because the choice of essence type for natural kinds is obvious. In this chapter, I will instead claim that several different essence types are scientifically respectable, and that the choice sometimes is *far* from obvious. In addition, I will argue that the choice of natural kinds and essence types are context-specific and depend on purposes of enquiry.

Before I can get to this point, I will first need to say something more about what I mean by natural kinds, and what I mean by essences and essence types.

In this thesis, I talk about essences as *defining* natural kinds, so that an essence specifies the necessary and sufficient conditions for belonging to that kind, and also gives explanatory power to natural kind terms.[227] My use of 'essential' and 'essentialism' is not the most common, and I will also apply these terms to areas where they are not so often applied (for example in biology and jewellery contexts). I claim that these terms are useful tools, and they have the advantage over the traditional conception of essence, which includes extra-theoretical continuity, of not being false.[228]

4.2 Natural Kinds and Natural Kind Terms

What *is* a natural kind term? This is not well-defined, and Ian Hacking advises against using the expression at all, because: "'Natural-kind term' is a devious phrase. It elides the distinction between the cosmic and the mundane."[229] What Hacking means is that there are two ideas behind the term, not necessarily yielding the same result. I will explore Hacking's suggestion that there are two separate starting points when explaining nature, with two correspondingly different ways to define what is a natural kind, but reformulate the starting points for my purposes. One of them is the terminology of natural languages (the "vernacular starting point") and the other the terms in scientific theories (the "scientific starting point").

For the vernacular starting point, the horses, roses, gold and pains we encounter are the phenomena we expect science to analyse and classify for us; they are the *explananda* of science. This starting point agrees well with a view that sees the role of science as explaining the world as we perceive it. Both Putnam and Kripke depart from here, as does David Chalmers, who builds an ontology from the vernacular starting point.[230]

---

[227] I will leave the term 'explanatory power' vague (similarly, Dupré talks about terms being "useful"). It is something that follows the ambition and purposes of individual sciences, something that is measurable, so that it makes sense to say that the explanatory power of one theory in a particular field can be greater than another. Putnam [1975], pp.295-296, gives us an indication of what is a powerful explanation and what is not with his example of why a square peg of a certain size will not go through a round hole. An explanation in terms of elementary particles and possible trajectories is a possibility, but it is a worse explanation than one based on geometry.

[228] See Dupré [1995] and Khalidi [2015] for convincing arguments against this stronger form of essentialism.

[229] Hacking [2010], p.291.

[230] See for example Chalmers [1997].

The vernacular starting point does not rule out an element of analysis, but that analysis is of a non-empirical kind. This is elaborated by the writers contributing to the so-called "Canberra Plan", where an initial step of conceptual analysis defines the area of search and success criteria to apply to subsequent scientific investigations. One of those philosophers, Frank Jackson, writes: "Our account sees conceptual analysis of K-hood as the business of saying when something counts as a K".[231]

For establishing conceptual continuity between terms in *scientific theories*, however, conceptual analysis is not enough as the first step. As Michael Ghiselin says: "[S]cientists do not attach a name to a class, then discover the defining properties which are its essence, but rather redefine our terms as knowledge advances."[232] For the purpose of analysing the development of scientific concepts, we need a model that can accommodate the dynamics when the definition of what there is to explain is affected by the results of the investigations.[233]

An alternative starting point for natural kind terms sees natural kinds as posits of scientific theories, generalised over in natural laws. According to this approach, natural kind terms become theoretical terms.[234] This is the position of W V O Quine. Being posits of scientific theories is the last word on the matter, because, as Quine writes about the existence of regularities involving natural kind terms: "[This] is an established fact of science; and we cannot ask better than that."[235] I will interpret this quote as a methodological principle saying that there can be no external principle overriding the choices made by scientific communities regarding natural kinds for their respective areas, and call it "Quine's Dictum".[236]

[231] Jackson [1998], p.46.
[232] Ghiselin [1987], p.135. Ghiselin's use of the term 'essentialism' is common, but not the one I just defined. I define an essence as the sufficient and necessary conditions determining a natural kind and giving it explanatory power. The role as a carrier of extra-theoretical continuity is an extra, alleged property of essences, and not a part of my definition.
[233] But I will complicate this picture in §4.3.
[234] In the following sense: "[A] theoretical term is one whose meaning becomes determined through the axioms of a scientific theory." Andreas [2017], first paragraph.
[235] Quine [1969], p.126.
[236] It is immaterial for me if Quine has exactly this interpretation in mind.

Most writers would agree that the practices and results of actual sciences are relevant to the philosophy of science, but I go further. I intend to take Quine's Dictum literally, use it as my default methodological rule, and draw the consequences. But it is a methodological principle, not a dogma, and I will from time to time discuss where it leads. In later chapters I will worry about whether Quine's Dictum has implications that are incompatible with realism. Even when doing so, I will assume that at least a weaker principle is sound, namely that any method proposed for the selection of natural kinds would justify many of the natural kinds referred to by existing sciences. I wrote in the first chapter that "I do not discuss whether scientific progress has actually taken place; that is the *explanandum*, what I in this thesis take as a given".[237] A project that wants to find an epistemic justification for scientific progress is not best served by a methodology that excludes a large part of its current posits. I will disqualify an approach that has such consequences, as it does not apply to *science* as we know it.

One method that risks excluding a large part of current sciences is the approach that seeks *a priori* rules for choosing natural kinds. It is influenced by the idea that there are genuine, objective kinds in nature, to be discovered by science. Other kinds might have their uses, but should be sharply distinguished from the natural ones, as *non-natural* kinds, or *artificial* kinds.

The question then arises how we could identify the genuine natural kinds. One line of thought is elaborated by Brian Ellis, who lists six *a priori* criteria that must be met by a candidate natural kind:[238] (1) Natural kinds must be mind-independent, (2) they must be categorically distinct, and (3) they must be demarcated from other kinds via intrinsic property. In addition, (4) members of the same kinds with different (non-acquired) intrinsic qualities must belong to different species of that kind, (5) memberships of two natural kinds cannot overlap, and (6) a natural kind – in contrast to other sorts of things – must have an essence that is a necessary and sufficient condition for belonging to the kind . I will not discuss these in detail, but only point to some problems Ellis has encountered.

[237] As stated in §1.5.
[238] In Ellis [2001], pp.19-21.

A priori criteria like those proposed by Ellis have the fundamental problem of being inconsistent with scientific practices. As Beebee and Sabbarton-Leary write, if we look at such practices,

> we would surely conclude that, for example, biological species are natural kinds, and we would therefore have no reason to expect categorical distinctness (criterion 2) to hold, and may have to abandon intrinsicness as well (criterion 3).[239]

In addition, Emma Tobin questions whether even chemical natural kinds meet the hierarchy requirement (criterion 5).[240] If we stick to Ellis, we risk eliminating most or all natural kinds in current sciences, violating my weak principle for their selection.

Magnus calls this approach, which holds that the posits of sciences should be evaluated against general criteria for proper natural kinds, "*a priori* philosophizing of the worst kind".[241] The opinion that science can be judged according to *a priori* criteria seems in need of a strong motivation. The motivation is often found in an alleged special status for the microstructural essence type, and I will discuss this in §4.6. below.[242]

According to the view defended, the view that follows Quine's Dictum, the kinds and essences postulated by a particular science are defined by that science and chosen for their explanatory force within the area of enquiry. If we take this view seriously, there are no *general* criteria for being a natural kind; the choice is up to each science. That does not make the traditional candidates, such as projectability/induction support, irrelevant – they are no doubt a good choice for most sciences – but it makes their role in choosing natural kinds indirect.[243] Instead, the choice of natural kinds ultimately depends on their usefulness for the science in question, that is how it best explains the phenomena in a science's domain, and "we cannot ask better than that".

[239] Beebee and Sabbarton-Leary [2011] p.3.
[240] Tobin [2010].
[241] Magnus [2012], p.20.
[242] This is in turn often motivated by a particular view of scientific realism. I will discuss this in the next chapter.
[243] For reasons why induction on its own is not a good *direct* criterion for being a natural kind, see Magnus [2012], pp.17-18.

Another Quine statement from the same article looks more problematic, namely when he says that the difference between the vernacular and the scientific is only a matter of gradation, because: "Science, after all, differs from common sense only in degree of methodological sophistication."[244] This quote might be read to underplay the leap sciences take from explanations relying on the functional essence type and observational properties, to the variety of essence types in use by current sciences, and the explanatory power that comes with this extension in tool set – and this would be a mistake.[245] What is true, though, is that with the approach I am defending, there is an absence of sharp boundaries between different types of human activities that aim to explain and predict.[246]

4.3 The Relationship between the Vernacular and the Scientific

Although I disagree with Quine that vernacular and scientific natural kind terms differ only in degree, they are nevertheless not independent; there is a complex, bi-directional influence between them. I will explore the relevant implications of that in this section.

I will start with how scientific kinds evolve from vernacular kinds using an example from chemistry. Kyle Stanford and Philip Kitcher discuss the term 'acid', which started out defined with criteria available to our senses.[247] I will regard this essence type as an instance of the *functional* type, in this case based on observational properties. The criteria for acid were collected by Robert Boyle in the 18th century. According to Boyle, acids have sour taste, are corrosive, change the colour of certain vegetable dyes, including litmus, and lose their acidity when they are combined with alkalis (bases).

Svante Arrhenius developed the first structural definition of 'acid' in 1887, suggesting that acids are hydrogen compounds dissolved in aqueous solution.[248] Johannes Nicolaus Brønsted and Thomas Martin Lowry (independently) proposed another definition in 1923, and

[244] Quine [1969], p.50.
[245] I will elaborate this in §4.7.
[246] See §4.6. I will therefore use a broad understanding of 'science'.
[247] Stanford and Kitcher [2000].
[248] That is a solution in which the solvent is water.

Gilbert Lewis put forward yet another in 1938. Boyle accepted the pre-scientific term as an unchanged starting point for 'acid', but this was not the case for the later scientists, who amended the term. Arrhenius and the others were looking for a common microstructure, and when doing so,

> Boyle's original list was modified because chemists hoped to use the properties attributed to acids to point to a common inner constitution and found that some of the phenomenological features of acids were more useful in doing so than others. Arrhenius, Brønsted and Lewis all drop the sour taste and corrosiveness requirements because they recognize that, at the molecular level, there isn't any common constituent of the structures that causally produce these kinds of features as well as the others on the list…the acid stereotype is modified in the course of chemical investigation so as to preserve a set of features that can be causally explained in terms of some common underlying structural property.[249]

For examples like this, the pre-scientific terms also show evidence of proto-scientific thinking: they are generalisations to achieve explanations and predictions, but in principle limited to what was naturally occurring and to observational criteria. But following what I suggest is *a change of essence type*, from a functional to a microstructural one, the definition of 'acid' was modified and its explanatory power increased.

The influence goes in the other direction too. The pre-scientific terms are by definition untouched by science; but our current vernacular terms are not, due to the status of science as a human activity that aims to find better explanations. Sometimes we just take the scientific terms onboard; we will not hold that whales are fish, even for non-biologists. In other cases, this is not practical.[250] As a result of this influence, also vernacular terminology needs temporal indices.

It is in light of these considerations that I can address a possible objection to my analysis: Did not the thought experiments constructed by Kripke and Putnam show that the rigidity of the

[249] Stanford & Kitcher [2000], pp.117-118.
[250] Dupré [1995], chapter 1, lists some examples.

terms and the necessity of their identification could be established by conceptual analysis alone, that is, established from what one would say in counterfactual situations? I am certainly not aware that I employ any essentialist principle when I follow Putnam and Kripke through these experiments, so it sounds implausible that it plays an operative role in my conclusion. But logically, the essence type premise is needed, as we saw in Chapter 2. The validity of the Twin Earth thought experiment not only relies on water consisting of H_2O across time and possible worlds, but also that this is the essence of water.

This issue can be resolved, I suggest, by considering conceptual development and the influence of science on our current everyday vernacular. It is part of our relation to science that we to a large extent are willing to be informed and corrected by scientific discovery. As a consequence of this, the identity of water with H_2O is implicit in the contemporary term 'water' and *part of its stereotype*. This is already built into 'water', so the thought experiment works. But it also *sets a limit* for thought experiments: they rely on linguistic intuitions and are bound by the temporal indices used, usually the 'Today' index, indicating that the scope of their conclusions is "given our language".

With this conclusion I can now also go back to two issues I discussed in the previous chapter. The first issue is a footnote where I said that it is doubtful whether Putnam's implicit assumption of an unchanged water stereotype over time is correct.[251] The increased knowledge within the community of chemical experts feeds into the normal language usage, into the vernacular vocabulary. The stereotype for water is not unaffected by scientific development; the stereotype for 'water' now contains its identity with H_2O. But this implies that the meaning of 'water' has *not* remained unchanged over millennia, as Putnam states it has, because for him, stereotypes are meaning components.

In the previous chapter I also left open the issue of what we can and cannot think. When I discussed Kripke's and Putnam's thought experiments, I wrote that "we cannot *truly* imagine that water is not H_2O, because if something is not H_2O, it is not water; H_2O is the essence of

[251] §3.2., footnote 6.

water."[252] Based on subsequent analysis, I now suggest that this argument is not best understood as an argument about our ability to think (imagine, conceive) in a narrow psychological sense. There is a better way to construe the argument, one that is not so obviously open to objections. We have again to take "our language" as a given, a language where our current vernacular natural kind terms have been influenced by their cousins used in science. The scientific natural kind *water* is an abstract object, defined by its essential property, H_2O. It is part of our common-sense attitude that sciences provide the more exact, deeper knowledge of reality. It is therefore reasonable to conclude that the abstract object that is water is also what the vernacular term 'water' refers to.

We construct thought experiments from our stock of current natural kind terms, with respect to both the subset of actual reality and the alternative scenario to which we compare it. We cannot when doing so, if we want to be consistent, enter an assumption that water is not H_2O *without changing the meaning of 'water'*.[253]

Having dealt with the possible objections to the dependency of Putnam's argument on essence types, I will now analyse this notion further.

4.4 Essence Types

I said earlier that Putnam's two argument for conceptual continuity both rely on an essence type assumption. I also said that "MoM" contains a defence of this assumption, namely that the choice of essence type is obvious. I will elaborate and question Putnam's defence in this chapter, after recalling Putnam's position.

In *N&N*, Kripke most of the time takes for granted that essences are microstructural. He writes, looking at a table: "[C]ould anything be this very object and not be composed of molecules? Certainly, there is some feeling that the answer to that must be 'no'."[254]

[252] §3.4.
[253] I will return to the issue of what we can and cannot think in Chapter 8, as it is part of Kripke's criticism of Physicalism.
[254] Kripke [1981], p.47.

Later on, he elaborates:

> [O]nce we know that this is a thing composed of molecules – that this is the very nature of the substance of which it is made – we can't then…imagine that this thing might have failed to have been composed of molecules.[255]

Putnam defends the same view in "MoM", adding his extra-theoretical ambition, in a series of thought experiments.

One is the story about gold and Archimedes. We saw in Chapter 3 that Putnam thinks that the extension of 'gold' has been unchanged since Archimedes' days (ignoring translation issues, as usual). Putnam holds that the $same_x$ relation for 'gold' was valid and operative for Archimedes as it is for us; it signifies a relation based on essential properties. Also the value of X has stayed the same, perhaps 'metal' in this case, certainly something that implies a "hidden structure" as its essence. We can say that, Putnam adds, because when Archimedes called a piece of metal "gold", he was at the same time making a statement about essence types, namely "that it had the same general *hidden structure* (the same 'essence', so to speak) as any normal piece of local gold."[256] In the previous chapter I used this example to show Putnam's dependence on a division of labour over time. But Putnam's main objective with the 'gold' example to establish a case against Kuhn's Challenge, defending extra-theoretical continuity, which Putnam believes implies the allocation of beliefs about essence types to historical individuals such as Archimedes. I will now discuss it from this angle.

I have already quoted the imaginary conversation between Putnam and Archimedes, where Putnam explains the difference in microstructure between gold and a sample of pyrite Archimedes believes to be gold, and where Putnam concludes by rhetorically asking: "[I]s there any doubt that he would have agreed with us that X isn't gold?"[257]

Putnam continues:

[255] Kripke [1981], p.127.
[256] Putnam [1975], p.235.
[257] Putnam [1975], p.238.

> If we had performed the experiments with Archimedes watching, he might not have known the theory, but he would have been able to check the empirical regularity that 'X behaves differently from the rest of the stuff I classify as *χρυσός* in several respects.' Eventually he would have concluded that 'X may not be gold.'[258]

Our intuitions in relation to the 'gold' thought experiment might be influenced by the slight differences in observational properties between pyrite and gold. We can imagine that Archimedes had perhaps noticed those differences and was looking for an explanation. But all this is irrelevant for Putnam's continuity arguments. If we introduce a variation to the experiment where we counterfactually imagine that all observational properties of gold and pyrite are identical, just as water_{E} and water_{TE} are in the Twin Earth experiment, Archimedes, when identifying a piece of pyrite as gold, would intend to use the term 'gold' in the same way as his peers, and succeed in doing so. But Putnam is committed to say that Archimedes still *would be wrong*, because the gold stuff is different from the pyrite stuff. The stuff is what matters; gold *must* have a microstructural essence. The microstructural essence must be what 'gold_{250BC}', not just 'gold_{Today}', relies on for its meaning, if Putnam is right.

Kim Sterelny, who defends Putnam against some of his critics, in effect illustrates why the assumption about unchanged essences (and therefore unchanged essence types) is needed; the semantic arguments do not get off the ground without them. Discussing an elaboration of the Twin Earth case put forward by Eddy Zemach, where someone from Earth *visited* Twin Earth in 1750, and naturally used the term 'water' to refer to samples of XYZ, Sterelny writes:

> Zemach has described a changed situation where the extension of 'water' includes both H_2O and XYZ. But that *is* a *change*. Water is H_2O. Our token of 'water' connects systematically with H_2O and no other substance…[259]

One interpretation of this quote is that Sterelny says that no instance of the vernacular natural kind term 'water', with any index, has referred to anything other than H_2O – but that is not

[258] Putnam [1975], p.237.
[259] Sterelny [1983], p.100.

strictly true. Zemach points out that historical uses of 'water' have included not only D_2O (still part of $water_{Today}$), but

> it is a historical fact that 'water' was regularly used to refer to a great variety of chemically dissimilar liquids, among which are tears, urine, sweat, saliva, solutions of ammonia and camphor, etc.[260]

Putting the details of the extension to one side, there is a more fundamental issue. Sterelny, like Putnam, takes the meaning index 'Today' as a given, exemplifying Kuhn's description of how scientists look back by describing history anachronistically from their own paradigm. This is often natural to do (recall Kripke's "given our language" proviso) but results so achieved cannot be used to draw conclusions about the relation of '$water_{Today}$' to 'water' with other indices. Sterelny has not shown that Zemach describes a change of '$water_{1750}$', only that this would be a change of '$water_{Today}$'. For the argument to extend to the former, we need to say that its meaning is the same as '$water_{Today}$', with an unchanged essence – and that requires an argument.

Putnam attributes a firm belief about essence types to Archimedes, but he does not assume that it was infallible. We could imagine a situation where Archimedes held such a belief about something that later turned out to *lack* a common hidden structure. "[T]he local water…may have two or more hidden structures – or so many that 'hidden structure' becomes irrelevant, and superficial characteristics[261] become the decisive ones."[262]

In my terminology, water (and gold) could have had a functional essence type based on observational properties. This fallibility of scientific identifications, which means that we can never know if the meaning of a natural kind term will change in the future,[263] seems to imply that we can never really know the meaning of a term. But this conclusion is avoided if we use the

[260] Zemach [1976], p.63.
[261] What I call "observational properties".
[262] Putnam [1975], p.241.
[263] It follows from the so-called "pessimistic meta-induction argument" that we should conclude that it *will* change. I will discuss this in Chapter 5.

temporal indices. We are well justified indeed to believe that '$water_{Today}$ is H_2O' is true. The issue is how we could justify the statement '$water_{Today}$ is $water_{Future}$'.

Putnam seems to regard the situation for 'gold' in Archimedes' days as analogous to the situation for 'water' in 1750, that is, Putnam thinks that the meaning of 'gold' has remained unchanged since Archimedes. The reason is that a microstructural essence type assumption was already "part of the original enterprise", to use Kripke's phrase. According to this reading, Putnam claims that the original enterprise in this case started *really* early, so that Archimedes when he talked about gold intended to refer to the stuff that had a sameness relation to a given sample in virtue of an underlying, hidden microstructure. But if such a view is problematic for the 1750 situation, where a hypothesis about such a microstructural essence exists, it is obviously even more so for the 3rd century BC, when Archimedes lived. Putnam here goes further than Kripke. His argument looks blatantly anachronistic, unless the choice of essence type really is – and always was – obvious. I will argue against that in the next section.

4.5 Alternative Essence Types

In the last section, we saw that Putnam insists that natural kinds such as gold must have a microstructural essence.[264] I said that putting beliefs about microstructure in the mouth of Archimedes requires a convincing argument. But maybe this is unproblematic, maybe microstructural essence types always are the obvious choice. This is certainly not how it looks. *Prima facie* there are least three major types of essences found in the history of science, defining three different types of natural kinds: structural, functional and historical. Structural essences can be microstructural (as in chemistry) or macrostructural (as for Linnaeus). Examples of functional essences include mental state types, if (one type of) Functionalism is right, and money in economics. Historical essences are found in biology (if Cladism is right)[265] and in individual persons (if Kripke is right). Quine's Dictum therefore leads us to recognise other essence types

[264] I will however in general not discuss individual essences, which is a mysterious subject, only essences assigned to natural kinds.

[265] Although often not called "essences" by philosophers of biology; a lack of microstructural consistency across classification often seen as an argument against essentialism. I discuss Cladism later in this section.

than microstructures on an equal footing as appropriate for scientific use. In order to make this point, I will look at two examples in the contemporary discussion where there are *rival* essence types, in the philosophy of mind/cognitive science and in biology. My suggestion in both cases is that Putnam's strong assumption about essence types is not generally correct; in these two cases the choice is far from obvious. I will in the next section also discuss some examples where a microstructural essence type assumption seems plainly wrong.

The Mind-Body Type-Type Identity theory ("the Identity Theory"), launched in the 1950s, looks for necessary and sufficient conditions for mental state types at a microstructural level, in the workings of the human brain.[266] Because an identity with neurophysiological states is supposed to be a scientific identification, in line with those in physics and chemistry, we can use my terminology to describe the Identity Theory as stating that being a certain type of neurophysiological state is the essential property of a particular type of mental state.[267] According to this theory, the explanation of what it is to be a particular mental state will eventually, at least in principle, be expressed in the language of neurophysiology. It is controversial which are the natural kinds to be so explained, where the majority of Identity Theorists favour an explanation from the vernacular starting point – that is, mental states such as beliefs, desires and sensations (perhaps with some modifications) – but a minority holds that the vernacular concepts are so polluted with bad philosophy that they must be eliminated.[268]

In the 1960s and 1970s, a number of articles and books by Putnam and by Jerry Fodor introduced the "Multiple Realizability Thesis" (MRT).[269] The MRT is based on the intuition that we might naturally assign mental states to creatures of very different physiological set-up, given similar enough behaviour. There appear to be interesting common features, for example regarding sensations like pain, that would not be affected by such differences; the mental states seem capable of being realized in multiple ways. Consequently, the Identity Theory could not be right, if interpreted as providing *necessary* conditions for being in a certain mental state. Brain

[266] See U.T. Place [1956] and J.J.C. Smart [1959].
[267] Original formulations often suffer from being expressed in terms of contingent identity.
[268] This is the view of Patricia and Paul Churchland.
[269] See Fodor [1974], Putnam [1967].

processes could give sufficient conditions for mental states, but necessary conditions and interesting explanations would have to be found somewhere else.[270] David Lewis puts it paradoxically: "Pain might not have been pain...Something that is not pain might have been pain."[271] But Lewis's point is not as absurd as it first appears. He can agree that on a token level the statement "this particular occurrence of pain is an occurrence of a particular c-fibre stimulation" is true, but contingently so: this particular occurrence of pain could have been identical (in a relevant sense) with a physical token of another type. I will come back to this argument in Chapter 8.

MRT is often combined with Functionalism in attempts to provide an account of mental states.[272] I said that the Identity Theory looks for necessary and sufficient conditions for mental state types in the workings of the human brain. Functionalism instead looks for functional definitions, independent of physical states. But my point here is not to argue for Functionalism; regardless of whether Functionalism of this type is an adequate description of the human mind, and irrespective of the complexities of MRT, there seems to be nothing wrong with the idea of functional essences as such. We should not rule out that Functionalism can be right just because it fails to postulate microstructural essences, when the theory is based on MRT that states that the right essences are to be found on *another* level. Functionalism cannot be wrong *a priori*. The intuitive force of MRT is exactly that the interesting stories about types of mental states – the stories with the greater explanatory power – cannot be expressed by neurophysiology, no matter how advanced, if it is possible for other types of creatures to be in pain. If we accept the existence of other essence types, Putnam's slogan "The meaning was different because the stuff was different" loses power to convince: the point about Functionalism, based on MRT, is of course that "the stuff" is *not* what makes a difference; the functional essence is.

[270] As applied to mental states, MRT comes in a radical and a less radical version. The latter allows for individuals with different physiologies (perhaps even with brain-equivalents based on silicon or green slime) to have mental states just like the ones we have. The former allows such differences even in a single individual, perhaps after a brain injury, or even during the activities of a normal brain.
[271] Lewis (1980), p 125.
[272] E.g. in Ned Block and Jerry Fodor [1972]. I will discuss Functionalism further in Chapter 8.

I now turn to the second example where the allocation of the best essence type is far from obvious: biological species.

What are the relevant similarities that make a group of animals into members of one particular species and not another one? There are many competing definitions of the term 'species' in biology, all with advantages and disadvantages. One type of definition is structural, either based on observable ("morphological") or on hidden structural features. But both morphological and genetic distinctions have a poor fit to the distinctions between the current life forms that we want explained. Macro-structural properties *within* a species often overlap the variation that exists *between* different species.[273] Regarding genetic criteria, Samir Okasha points out that there for example is no genetic, microstructural property that all chimpanzees, and only chimpanzees, have.[274] Okasha concludes:

> Empirically, it simply is not true that the groups of organisms that working biologists treat as con-specific share a set of common morphological, physiological or genetic traits which set them off from other species.[275]

The situation is even more problematic with a Darwinian perspective of evolution over time, or for counterfactual scenarios. Peter Godfrey-Smith writes:

> [S]uppose we identify a set of distinctive genetic features of (say) our own species, *Homo sapiens*. Would it be *impossible* for an organism to live a recognizable human life without these genetic features?[276]

As Godfrey-Smith says, we do not know the empirical answer to this question, but the question certainly seems open. If the answer *could* be "no", this suggests that the genetic features in

[273] Dupré [1995], p.54.
[274] Stanford and Kitcher [2000], pp.120-121: "An attempt to fix the reference by declaring chimpanzees to be those organisms whose somatic cells carry chromosome pairs with a specific banding pattern and with a certain arrangement of special loci would be hopeless, since chimpanzees, like other mammals, can have abnormal karyotypes (trisomies, for example), significant deletions, translocations, and all the usual genomic disruptions that afflict their evolutionary cousins (namely, ourselves)."
[275] Okasha [2002], p.196.
[276] Godfrey-Smith [2014], p.106.

question, would we find them, might not be essential; they would not explain what it is to be a human being.

The evolutionary perspective is of central importance for modern biology, and definitions of 'species' in terms of change over time ("phylogenetic") have therefore become increasingly popular.[277] Phylogenetic definitions utilise ancestor and descendent relations rather than internal properties. Cladism is the most widely-accepted of the phylogenetic theories, and I will use Cladism as the example.[278] Other criteria, including observable features, are heavily used also by phylogenetic theorists, but only to form phylogenetic hypotheses; they do not *define* taxa such as species. This an example of a historical essence type, as it is the historical relations that give explanatory power when analysing what a species is, not the "stuff" they are made of or their observational features. On an "evolutionary tree", which is an often-used metaphor, all life forms that have existed on earth are represented on branches, with the earlier forms closer to the bottom, and the species pictured as the segments between branching points.

There have been objections to historical essences.[279] One objection is that they rely on external properties, in the sense of relations between entities. But this objection is saying little else than that the writer raising it prefers structural essences, which is what cladists are rejecting. A second objection is that phylogenetic essences do not guarantee extra-theoretical constancy. This ambition is indeed often a part of traditional essentialism, but it is not a part of the type of essentialism that I employ in this thesis. For me, extra-theoretical continuity is an additional hypothesis, not a part of the definition.[280] Dupré notes that in a phylogenetic system such as Cladism, unless everything living thing is part of one common taxon, there is a need to pick a common ancestor for each species as the place to start, and no general rules for this come with

[277] This is however still controversial, and some philosophers of science still hope for a microstructural essence to be found. Eileen Walker [2012], p.151, writes: "Putnam and Kripke are assuming that the DNA of an organism can be used to identify the species to which it belongs – however we choose to name or define that species. Despite the consensus against this view, I am suggesting that they were right after all."
[278] A "clade" is defined as an ancestral group and all of its descendants.
[279] See for example Devitt [2008].
[280] I define an essence as a property that "specifies the necessary and sufficient conditions for belonging to that kind, and also gives explanatory power for other properties".

the theory. "Phylogeny, in short, cannot possible create essences *ex nihilo*."[281] This gives a good reason to be suspicious of the claim that essences are extra-theoretical, but not of the notion of historical essences as such, understood without extra-theoretical claims.

My aim is no more to defend Cladism than it is to defend Functionalism; so far, I have only argued that the choice of essence type is not always trivial. If we follow Quine's Dictum and (tentatively) accept the natural kinds as those posited by and generalised over by current sciences, these two theories look like viable candidates – unless there are additional reasons to exclude them from the competition. And if there are several essence types in contention (if there are different essence types usefully employed by sciences), this means that it would not help Putnam if he was right regarding Archimedes and gold; the example would not serve as a general pattern, applicable to other cases.[282]

Furthermore, it is not obvious that the conclusions Putnam draws from the 'gold' thought experiment are the only ones possible. Putnam introduces a pre-mature scientific crisis by exposing Archimedes to empirical findings (the different behaviour between pyrite and gold) and also a new theory (modern chemistry) that better explains the data. If Archimedes represents the best available, relevant knowledge at the time, can we then conclude that he and everybody in his position who received the information from Putnam would embrace the new theory? No, we cannot. For situations we know a bit more about, Kuhn shows that some practitioners under the old paradigm move over to the new paradigm, while others stick to their old ways until they die. The historical changes Kuhn describes are rather messy affairs, where it often is unclear who actually presented the new paradigm and when it was accepted. New ideas might also at least for a period be expressed in the terminology of the old theory; Cavendish and Priestley arguably refer to oxygen when they use the term 'dephlogisticated air'.

[281] Dupré [1995], p.57.
[282] I will discuss some reasons advanced to exclude all non-microstructural essence types from the definition of true natural kinds in the next section.

It is of course possible that Putnam might have convinced Archimedes in the way he describes the encounter. But for Putnam's case, we would also need to assume that Archimedes believed in microstructural essences *before* the encounter. If the outcome was that he agreed with Putnam, this could be seen as an example of continuity of Archimedes' old beliefs, with some refinement achieved by the discussion. This is Putnam's intended conclusion. But it could just as naturally be described as a conversion of Archimedes to the new theory, where he *gives up* his old beliefs in the face of Putnam's convincing arguments, and this would represent change rather than continuity. The issue is not whether modern chemistry is more powerful than previous theories, but whether there is continuity between Archimedes' usage of 'gold' and later uses of the term.

In addition, could Archimedes not, faced with Putnam's empirical evidence, instead have concluded that gold can have two different microstructures? In Chapter 6, I will discuss examples when something very similar to this in fact happened, and I will endorse Joseph LaPorte's view of the crucial role of decision-making. If Putnam is right in his example about Archimedes and 'gold', he is not trivially right – and therefore not *generally* right. I will also, in Chapter 5, discuss the reason that Putnam in "MoM" feels that he *must* be right, and what the alternatives otherwise are.

Before we get there, I will in the next section discuss whether there are reasons to believe that microstructural essences should have a special, favoured status, making a choice of such essences a rational goal.

4.6 The Special Status of Microstructure

As we have seen in earlier chapters, it is sometimes suggested, usually with examples from physics or chemistry, that microstructural essences are the only *real* essences, and natural kinds defined by such essences the only real natural kinds. Others are at best temporary place-holders. When writing "MoM", Putnam believes that essences are microstructural, and that this is obvious, the only alternative.[283] We can therefore without hesitation assign the relevant beliefs and intentions about water to people in 1150 and about gold to Archimedes in Ancient Greece.

[283] At least essences outside the human mind; Putnam defends the MRT.

Paul Churchland draws the consequences of this when he states that

> the only genuine natural kinds appear to be those comprehended by absolutely the most basic laws of our science. On the view here outlined, mass, length, duration, charge, colour, energy, momentum, and so forth all turn up safely as natural kinds or properties. But precious little else does.[284]

It does not follow from my definition of 'essence' that the microstructural essence type is the only serious contender; according to my definition, essences provide necessary and sufficient condition plus explanatory power. Nor is it implied by scientific practices; it is at odds with Quine's Dictum. I have already mentioned that the requirement that real essences should be microstructural sits badly with current biology. But this is not specific to biology. Muhammad Ali Khalidi's main example comes from a branch of physics, namely fluid mechanics.

> [This] is a macrolevel science at least some of whose properties and kinds simply have no counterparts at the microlevel (e.g., the property *viscosity* and the kind *Newtonian fluid*) and are not properties and kinds of atoms and molecules (much less elementary particles).[285]

There seem to be many perfectly valid examples of other types of essences, so what reason would there be to deny them the status of being essences, violating Quine's Dictum? Those who insist that only microstructural essences can be accepted, whom Khalidi calls "microphysical fundamentalists",[286] must convince us that these scientific practices somehow are flawed. I will discuss three types of arguments for this: an argument from physical constitution, an argument from causality and a related argument from natural laws.

One reason to favour the micro over the macro is the argument that the micro entities *constitute* the macro entities, so that different macro level entities could be formed by the same basic ones: a table is "nothing-but" its constituent particles. But this is a very far-reaching argument, since it

[284] Paul Churchland [1985], pp.12-13.
[285] Khalidi [2015], p.84.
[286] Khalidi [2015], p.39.

seems to deny real existence to everything except the lowest level of matter. As Khalidi points out, it also relies on there *being* a lowest level. If there is not such a level, would it follow that there are no natural kinds? Or if we can never conclusively establish whether this is the case or not, which seems to be a realistic possibility,[287] would we never know if there are natural kinds or not, let alone which they are? Paul Churchland is prepared to live with this. He writes:

> [It is] a wholly empirical question whether or not the universe is…like an 'explanatory onion' with an infinite number of concentric explanatory skins. If it is like this, then there are no basic or ultimate laws to which all successful investigators must inevitably be led, and…there are no natural kinds.[288]

But natural kinds have now become elusive, unavailable for practical work in the philosophy of science, and in breach of my weaker methodological rule that a selection principle must not rule out a large part of current science.

A second reason to favour the micro over the macro could be a worry about causality. For methodological reasons,[289] causes should be located on a basic level only, it is argued, and causal effectiveness is what matters for laws of nature, which generalise over natural kinds. It is often added that these laws of nature must be *exceptionless*, which is supposed to be true on a low level and not true in higher level sciences.

The unpleasant conclusion about causality just discussed is avoidable, however, as there are alternative accounts of causality available, accounts that do not insist on ruling out all these *prima facie* causal explanations. According to Nancy Cartwright, there are a variety of different kinds of causal laws, related to different types of causal questions. These laws, she suggests, share common properties, but "there are no interesting features that they all share in common."[290] James Woodward, similarly, favours

[287] Khalidi, [2015], p.38: "It would take more energy than is currently available, or indeed may ever be available, to conduct the scattering experiments needed to determine whether quarks have inner structure. It may be that there are further levels of structure at yet smaller scales, which we will not, and perhaps cannot, uncover."
[288] Paul Churchland [1985], p.14.
[289] One such reason is the perceived risk of over-determination.
[290] Cartwright [2004], p.814.

> a broad notion of causal explanation according to which, roughly, any explanation that proceeds by showing how an outcome depends (where the dependence in question is not logical or conceptual) on other variables or factors counts as causal.[291]

Perhaps the causal-looking regularities fail to be causal laws by some standards. If so, we can give them another name, or as a general rule accept that they are *causal-looking* explanations, featuring in *law-like* generalisations; placeholders until a future collection of superior sciences does away with them all. However, in the meantime, they are what the best sciences offer providing obvious explanatory value.

I also mentioned the third, proposed reason to favour microstructural essences, the requirement that proper natural laws should be exceptionless. This again clashes with existing practices. There are plenty of examples in the special sciences of causal-looking regularities that have explanatory power, without being exceptionless. In economics, the law of supply and demand says that when demand increases and supply is held fixed, price increases. On the face of it, this is a (law-like) regularity expressing a causal relationship. It certainly does not hold in all conceivable circumstances (not in regulated markets, not when the sellers do not have the relevant information, etc.) but it is not obvious that this is a fatal problem.

In addition, similar things can be said for the laws of physics. Cartwright writes: "[T]here are no exceptionless quantitative laws in physics. Indeed not only are there no exceptionless laws, but in fact our best candidates are known to fail."[292]

Khalidi makes the same point. The exceptions, he claims,

> are often due to interactions that can best be explained at another level of description. In the case of the general statement that viscosity decreases with an increase in temperature in liquids, one exception involves the element sulfur, whose viscosity increases at a certain temperature between its melting and boiling points because polymerization occurs

[291] Woodward [2003], p.6.
[292] Cartwright [1983], p.46.

> and there is a change of allotrope. This exception can be explained by referring to microlevel reality.[293]

Examples, Khalidi argues, can be found also on a sub-atomic level. In the end we are again left to speculate that a lowest level, if it exists, might remove the problem.

> But perhaps it may be said that the reason for there being exceptions in this case is that this is not the most fundamental level and that if we really descend to the level of quarks and leptons (or whatever the bottom level turns out to be, if there is one), then we would find truly exceptionless laws. [294]

But not even quark-level laws would necessarily be exceptionless, due to potential impact of quantum effects, Khalidi adds.

Khalidi draws the conclusion that being exceptionless cannot be a criterion to differentiate basic and special sciences. He adds that the more frequent occurrence of exceptions in some sciences can be plausibly explained by the systems in such sciences being larger and more complex, and therefore more exposed to interference from other systems.

I will add an example of a causal-looking explanation in sociology.[295] In the early 20th century, Robert Michels described a regularity by which trade unions and political parties over time gradually become more conservative and less democratic. H. Richard Niebuhr later showed that this regularity also transcended the original area of application by successfully applying it to religious organisations, increasing its generality. If this regularity is correct, it can hardly be so without exceptions and qualifications – and nobody expects this either. Nevertheless, the law-like regularity identified by Michels and Niebuhr might say something important about at least a certain type of organisation, and we could look for common mechanisms that could explain it.

[293] Khalidi [2015], pp.105-106.
[294] Khalidi [2015], p.106.
[295] Quoted in Bruce [1999], pp.50-55. More recent examples suggest themselves.

I will not enter too deeply into the complexities of causality and natural laws, but I will note that the consequences again are so drastic that these argument approaches a *reductio ad absurdum*. They assume limitations on science that would rule that a very large number of theories with good explanatory power are not proper sciences, falling foul of my weak methodological principle. This suggests an over-reliance on physics as a role model for what a science should look like, and a possible over-confidence as to the clarity of causation and natural laws.

There is also a more principled argument against the idea of a reduction to "basic sciences", namely that causal (or causal-looking) explanations are sensitive to their theoretical context. We have encountered one argument of that type, namely MRT, but examples are by no means restricted to the philosophy of mind. Consider again the law of supply and demand: there have to be suppliers, goods/services and buyers for any market to function, but the specific nature of these is entirely immaterial for the purposes of this regularity, and such details would add nothing at all to its validity. I will later talk about the context-sensitivity of natural kinds in terms of purposes of enquiry, adding more examples.

In §4.2 I introduced two approaches to the selection of natural kinds in science: Quine's Dictum and *a priori* rules. In this section, I have discussed some potential reasons to reject Quine's Dictum and instead apply philosophical *a priori* rules to regulate what would count as proper natural kinds. None of these arguments are totally convincing, and all of them have drastic consequences, as they imply that many or perhaps all of the natural kinds postulated by current sciences (and by extrapolation also future sciences) fail to qualify as proper natural kinds. Instead of a theory explaining the nature of our natural kinds, which looks like a legitimate and important goal, we have ontological positions that *deprive* us of kinds that we can use, replacing them with natural kinds in an ideal science we might never know.[296] Many writers are uneasy about this direction, and Magnus writes: "Natural kinds, so defined, would be the abstruse promise of a hoped-for future science. We would still want a term for the categories apt to actual

[296] Magnus [2012], pp.21-22: "One may retain [microstructural] fundamentalism only at the expense of current science".

science."[297] This agrees with my weak methodological rule of thumb, and I will reject "microstructural fundamentalism" for this reason.

If we instead apply Quine's Dictum, we can with Magnus continue to use the term 'natural kinds' for the kinds postulated by current sciences (and science-like exercises). We can also continue to hold the natural thought that there are several types of essences, and that these are picked by each science for reasons of explanatory power. If we take the classification systems of a science at face value, and accept that theories, in a fallible way, justify their postulates of natural kinds, we must also accept that different essence types are employed across scientific communities.

During this analysis, we came across the idea of essences being dependent on their theoretical context and the purposes of the enterprise for which they are postulated, be they the purposes of economists, jewellers, chemists or bakers.[298] The next section will explore that idea.

4.7 Natural Kinds and Their Context

I have so far talked about "essence types" for two reasons: (i) to point to the need for a theoretical context before essences of natural kinds can be identified/assigned, and (ii) to open up the possibility of alternatives to the microstructural type sometimes assumed to be the one and only essence type. I also said that the assumption of the primacy of microstructural essences has a connection with a view of realism, which I will discuss in the next chapter. We found that such an insistence on microstructural essences clashes with existing scientific practices. I will now discuss what these practices instead suggest, and in particular the idea that the choice of natural kinds and their essences depends on the issue at hand and the need for natural kinds with explanatory power.

One suggestion is put forward by Khalidi. Arguing against simpler models of how sciences relate, he proposes that we should talk about different *domains*, separated by differences in the

[297] Magnus [2012], p.22.
[298] Magnus [2012], pp.133-136, describes the last example.

size and duration of their natural kinds, and by the objectives a science is trying to meet.[299] Similarly, Magnus argues that natural kinds are legitimate and useful, but that they are always specific to a *domain of enquiry*.[300] "If I say 'Water is natural kind', then I have said something semantically incomplete. Naturalness is a two-place relation…[W]ater is a natural kind for the domain of chemistry."[301] If Magnus is right, the heterogenous appearance of current science is not a fault, but a strength, increasing the overall explanatory power of science. We have jade as a natural kind for jewellers, but jadeite and nephrite for chemists. Admittedly, jewellery is not usually thought of as a mature science, but it is a consequence of the present line of thought that there is no sharp border between mature sciences and other human endeavours. There are generalisations regarding the manufacturing, sales and decoration connected with jade, supported by treating it as a natural kind, with a functional type of essence.[302]

One of Magnus's examples is planets, needed for generalisations in astronomy,[303] but with no specific or unique chemical or physical properties in common. Planets are functionally defined, in terms of their relations to other celestial bodies, and they are useful when explaining things like solar systems.

Cartwright also stresses the role of purposes when she writes about practices in physics:

> We construct different models for different purposes, with different equations to describe them. Which is the right model, which the 'true' set of equations? The question is a mistake. One model brings out some aspects of the phenomenon; a different model brings out others…No single model serves all purposes best.[304,305]

[299] Khalidi [2015].
[300] Domains of enquiry do not need to be discrete, but can overlap.
[301] Magnus [2012], pp.42-43.
[302] Richard Boyd [1999], p.148, has expressed similar views. "It is widely recognized that the naturalness of a natural kind – its suitability for explanation and induction – is discipline relative." [1999, p.148]
[303] See Magnus [2012], pp.76-77, for an example.
[304] Cartwright [1983], p.11.
[305] For Cartwright, models are postulates are more ontologically stable than fundamental laws, and their success dependent on how well they approximate phenomenological laws. (The book I am quoting is called *How the Laws of Physics Lie*). She adds, in Cartwright [1983], p.17: "There are always more phenomenological laws to be had, and they can be approximated in better and in different ways. There is no single explanation which is the right one". I will discuss scientific models being dependent on purposes of enquiry in Chapter 5.

I will stay close to Cartwright's terminology and talk of natural kinds as being sensitive to the *purpose of enquiry*.

Natural kinds have their role in a certain context, and they draw on other pieces in this context to give them their explanatory role. Magnus writes: "the kind **gold** only supports inductions in the context of chemistry"[306] mentioning its disposition to dissolve in *aqua regia.*[307] We can perhaps amend and expand his example a bit by saying that the chemical natural kind gold supports *certain* inductions in the context of chemistry. But gold is also a natural kind, supporting *other* inductions, in a foreign exchange context. As such, it is useful if the purpose of the enquiry is to examine the rise and fall of the Bretton Woods system. But while that type of enquiry posits the same metal, gold, there is no need for details about the chemical composition.

Making natural kinds enquiry-specific is not trivializing them, Magnus says. According to him, something is a natural kind for a domain if it is indispensable for carrying out successful scientific activities in that domain, and a successful science is one that makes sense of, explains and predicts the phenomena within its scope.

Attempts to define robust features of any kind across all sciences are put in doubt by the existence of multiple essence types and multiple (and incompatible) perspectives between and within sciences. This is the lay of the land. The question is whether this is a problem – in particular if it is a problem for realism. I turn to this question in the next chapter.

4.8 Conclusions

I have distinguished two starting points for natural kind terms, the vernacular and the scientific. For my purposes, the justification of scientific progress, I focus on the latter. I have chosen Quine's Dictum as my methodological rule and rejected the alternative approach, the search for *a priori* rules.

[306] Magnus [2012], p.41.
[307] A mixture of nitric acid and hydrochloric acid that can dissolve e.g. gold and platinum.

I have found that Putnam's two ways to respond to Kuhn's Challenge both rely on an essentialist assumption that says that the essence type must have been constant through history. When writing "MoM", Putnam assumes that essences are microstructural, and that this is obvious, the only real alternative. I have wanted to complicate this story by pointing to cases where the choice of essence types is not obvious at all. I do for example not need to assume that Cladism is correct, because if Cladism is an alternative to be taken seriously, Putnam is already in trouble, since in that case the choice is *non-trivial*. This would mean that the assignment of beliefs and intentions concerning microstructure to the long dead becomes problematic and therefore in need of an argument. Similarly, it seems neither confused, inconsistent or obviously wrong to believe that there is an identity relation between mental states and particular physical structures on a *token* level, but that the identity on a *type* level is between mental states and functional states. Furthermore, if we entertain the not-implausible idea that the essential properties for the people living in Archimedes' days might have been the functional ones, and that the abstract object defining water as H_2O or gold as the element with atomic number 79 came with modern chemistry, that would be inconsistent with Putnam's claims. In addition, there are areas where microstructure have no explanatory power; I mentioned the definition of planets and the law of supply and demand.

Insisting on microstructural essences as the only ones proper for natural kinds clashes with scientific practices, and consequently clashes with Quine's Dictum. I have discussed some reasons to believe that only the microstructural essences are the true essences, and only the natural kinds so defined are the *true* natural kinds; but their application would rule out a large part, if not all, of the natural kinds in current sciences. Indeed, this seems to be a risk for any general *a priori* set of principles. Defending an acceptance of existing practices, I placed the choice of natural kinds and their essences (and essence types) within purposes of enquiry, contexts in which posits are chosen for their explanatory power, for a specific purpose.

Why would anyone object to this path? I believe there is a more fundamental reason than insistence on *a priori* rules, a deeper motivation for the insistence on a common type of natural kinds and essences. This reason is a certain view of sciences, motivated ultimately by a particular

type of realism. I address this issue as part of my discussion of the “Perfect Theory Theory” in the next chapter, where I also present an alternative, which is compatible with multiple essence types and agrees with Quine’s Dictum.

5 Scientific Realism

5.1 Introduction

There are several versions of realism, putting forward metaphysical, semantic or epistemic arguments (usually combined), but they have a common starting point in a metaphysical thesis: the world exists independently from the human mind, and independently from human inventions. This basic metaphysical thesis is a pre-requisite for all types of realism. Basic epistemological realism adds that it is possible for us to acquire knowledge of this world. I will treat these types of realism as confirmed enough by everyday experiences and evolutionary arguments, and not question them here.

Scientific realism claims that scientific theories which postulate natural kinds and regularities can provide such knowledge. It is scientific realism that I will focus on in this chapter when I discuss Kuhn's Challenge that consists of two questions:

> (A) *If the meaning of key terms change between theories on either side of a paradigm shift, how can we say that these theories are about the same thing?* And,
>
> (B) *Even if we assume that two theories do address the same subject matter, how can we determine which one is better?*

Intuitively, we think of scientific practices as guided by a wish to better understand and control this objectively existing reality. This is indeed a reason that Kuhn fails to convince in his section X argument in *Structure*, where he moves from statements of how the world appears to scientists, to a conclusion about how the world is.[308] The realist intuition needs to be accounted for. For scientific realism, the metaphysical discussion has often concentrated on non-observables posits. It is one thing (and not usually controversial) to believe that desks and tigers

[308] See Chapter 1 in this thesis.

exist,[309] and another to believe that electron and quarks exist in the same sense; the difference is that theories have postulated different non-observables, or changed the defining characteristics of their non-observables, during the history of science. But while there are many versions of scientific realism, a whole-hearted realist about science believes that our best scientific theories include true or approximately true descriptions of both observable and unobservable entities in an objectively existing world. Because realism stresses the independence of the world from us, the scientific realist regards theories as fallible; we can always make mistakes or be unable to discover features of the world. The realist indeed regards many statements of previous theories as false, and many of their posits as non-existent.

The combination of descriptionism and Kuhn's theory of paradigm shifts invites anti-realist conclusions, and we saw that Kuhn – on and off – is tempted to draw such conclusions. If reference, as descriptionists believe, is determined by descriptions constituting the meaning of a natural kind term; and if this meaning, as Kuhn argues, might change with paradigm shifts; both the justification of scientific progress and scientific realism are in question. In particular the epistemological thesis for scientific posits seems to be in trouble, and a mind-independent world is not much use to science if we cannot learn anything about it.

Larry Laudan, who is not a realist, expresses his realist intuition in this way: "All of us would like realism to be true; we would like to think that science works because it has got a grip on how things really are. But such claims have yet to be made out."[310] I will call proposals to "make out" claims regarding scientific realism – to give us acceptable reasons to *believe* in scientific realism – "justifications". Proposed justifications, I suggest, will need to meet four criteria:

(i) The justification must *cover natural kinds and individuals*. These are the entities that my discussion about Kuhn's Challenge focusses on. I will rarely discuss regularities and laws, although there is a close connection: natural kinds are those entities a science formulates regularities and laws for.

[309] However, Peter van Inwagen [1990] accepts tigers, but not desks in his ontology.
[310] Laudan [1981], p.48.

(ii) The justification must *be available to us*, at least in principle. There might well be objects and events we are unable to discover, but nothing can count as a justification if it is forever beyond our reach.

(iii) It must *make scientific theories fallible*. As I said, the world is, for a realist, independent of us, and we are undeniably prone to mistakes. Our theories can therefore always be proven wrong. To hold that our theories could capture all aspects of reality in any given field would be arbitrary (why *our* theories?), and as Khalidi puts it, "misguidedly anthropocentric".[311] He concludes,

> The claim that scientific categories correspond to natural kinds need not imply that *all* natural kinds will be successfully enumerated, even at the end of (human) inquiry…Realists about natural kinds may need to content themselves with the *truth* and *nothing but the truth* without also insisting on the *whole truth*.[312]

(iv) It must *sort historical posits* in the right way, separating referring historical terms from non-referring ones. Stathis Psillos makes this point.

> In order for realists to defend the claim that there is some substantive continuity in revolutionary theory-change, they have to show that not all abandoned theoretical terms are in the same boat as 'phlogiston'…[313]

There is a close connection between my main subject at hand, the issue of scientific continuity, and the issue of realism, because scientific realists usually see scientific progress as an increasingly better description of this reality. Realism typically assumes the conceptual continuity of natural kind terms, but only for *some* of these terms. Because Kuhn's doubt over the criteria for continuity invites anti-realist conclusions, many realists, such as Stathis Psillos, have pointed to a dependence of realism on conceptual continuity across changes in scientific theories. I will support Psillos's point by discussing, and rejecting, the possibility of realism

[311] Khalidi [2015], p.219.
[312] Khalidi [2015], p.219.
[313] Psillos [1999], p.292.

without a dependence on conceptual continuity in §5.4.1. In other words, full justification for scientific realism also needs an account of conceptual continuity.

In this chapter, I will discuss several arguments defending either a fully-fledged or a partial scientific realism. One influential view that plays a central role in the discussion about both continuity and scientific realism is what I call "the perfect theory theory" (PTT). This refers to a family of related ideas and assumptions that offer a solution for continuity and a justification for scientific realism. Despite its well-known weaknesses, the PTT has often been regarded as necessary for scientific realism, and many arguments for realism implicitly rely on some of its assumptions. But I will argue that the PTT is untenable and therefore these arguments fail. Instead, I will defend a realism *without* the PTT.

In my analysis, I will rely on a methodological principle I introduced in the last chapter and called "Quine's Dictum". It says that we shall recognise the practices of our current sciences, because "we cannot ask better than that". However, I also said that will discuss its implications. In this chapter I will discuss a case where its application might look like a threat to realism: the heterogenous nature of current scientific practices. I will conclude that it is not.

5.2 Truth and the Perfect Theory Theory

Metaphysical realism claims that the world exists independently from the human mind, and epistemological realism that it is possible for us to acquire knowledge of this world. Scientific realism states that scientific theories can contain such knowledge. On top of this, the "Perfect Theory Theory" (PTT) adds the idea that there in principle could be a *perfect* theory, where all natural kinds and regularities postulated have a one-to-one relationship with features in the world.[314] As it is often expressed, the perfect theory would "cut nature at the joints".[315] This version of scientific realism makes a claim not only about the world as such, but also about its relation to our *theories*: the world is such that our theories can capture its true features and distinctions.

[314] One influential example I will discuss is Putnam's realism in "MoM", but the ideas are widespread.
[315] A phrase going back to Plato's *Phaedrus*.

All statements we derive from the perfect theory will naturally be stable; they describe the objectively existing world, and there is a one-to-one relation between natural kinds in the world and natural kind terms in the perfect theory. We are entitled to call this theory as a whole "the true theory", as it perfectly matches features in the world, and all its derived statements are true.

Psillos endorses this idea when he envisages that the world has "a definite and mind-independent natural-kind structure."[316] The choice of words, "natural-kind structure", suggests an assumption about theories and about our ability to obtain knowledge of nature's joints, discovering nature's secrets.

An endorsement of PTT in the form discussed so far naturally leads to the following view of scientific activities:

> $PTT_{Methodology}$: Science should seek to formulate the perfect theory, by discovering nature's joints.

It also implies a quality criterion for scientific theories, by which they can be compared:

> $PTT_{Quality}$: A theory is better the more of the perfect theory it duplicates, the more truths it contains.

I will call a version of the PTT that proposes these two points "Absolute-PTT" as a contrast to the version discussed in the next section.

The Absolute-PTT gives an answer to both questions (A) and (B). It has an answer to (A) because not only do the natural kind terms in the perfect theory correspond totally to objectively existing natural kinds, our current scientific theories capture some of that, and continuity is guaranteed by $theory_{Later}$ picking up those elements from $theory_{Former}$, complementing them with new elements. This is the basis that makes it possible to talk about the perfect theory as a solution to question (B) too: given that we (in principle) have an uncontroversial identification of natural kinds, we can talk about a convergence of science towards the perfect theory if the truths

[316] Psillos [1999], p.xix.

of older theories are kept, perhaps as special ("limiting") cases, and new truths are added by later theories. A theory is better than another if it resembles the perfect theory more.

The Absolute-PTT implies the existence in principle of *ex ante* criteria for continuity, where all data needed about the next step forward can be available before the event, since the objectively existing natural kinds have properties that must be described in the correct way. $PTT_{Methodology}$ tells us that this is what scientists should aim for and $PTT_{Quality}$ how they should measure their success. A central feature of the perfect theory is that it is *one***;** if there is one world, there is one perfect, integrated theory mirroring it.[317] Consequently, all different scientific theories must fit together, if they get it right. This leads to a requirement for a *common denominator*. Higher-level sciences must be able to reduce to basic sciences if they are to be true descriptions of the world.

Paul Churchland gives a clear formulation of this view, which he regards as problematic and subject to empirical confirmation.

> [T]here exists some final, uniquely true theory whose laws express the basic regularities in the universe and whose predicates denote its most basic kinds. It was not supposed that we will ever possess such a Utopian theory: only that our currently best theories give us our current best shot at reality's basic laws and basic kinds. And that basic laws and kinds *are there* to be aimed at.[318]

The perfect theory is perhaps there from a God's Eye point of view, but not from ours. Churchland calls it "utopian". For a realist, the world exists independently from us, and due to limitations in our cognitive abilities, time or physical resources, we will naturally often be unable to unearth all aspects of reality. We can obtain knowledge about the world, but what we think we know (possibly with some *a priori* exceptions) will always be fallible. And herein lies an obvious problem for this notion.

[317] This is sometimes referred to as the idea of a "unified science".
[318] Paul Churchland [1985], p.14.

The Absolute-PTT might be elegant and powerful in principle, but not in practice, and it can therefore not be an acceptable justification for realism. If we stumbled across a piece of the perfect theory in some area, we would not know that we did, for scientific theories cannot be conclusively proven. The Absolute-PTT might give a neat definition of the relative value of theories in $PTT_{Quality}$, but not one that we can use. Nor can we can follow $PTT_{Methodology}$. Kuhn draws attention to the requirement of access to the perfect theory for any practical application: to be able to compare two theories in respect to their proximity to (a realist's notion of) truth, we need to have a look at that perfect theory too; because we cannot, the method is not useful.[319] Consequently, new paradigms are chosen for a variety of reasons, but getting increasingly closer to the perfect, true theory is not one of them.[320] The answer Absolute-PTT gives to questions (A) and (B) of Kuhn's Challenge cannot work, because the perfect theory is not available to us. It fails criterion **(ii)**.

Putnam comes to the same conclusion when he writes "Three Kinds of Scientific Realism"[321] a few years after "MoM". In this article, Putnam uses an epistemological argument against the PTT,[322] namely the argument that different descriptions can fit the same facts: there can be two mathematically and empirically equivalent theories that nevertheless are incompatible. A realist who insists that one of the two theories must be the *true* one, the one corresponding to reality, "pretends to a notion of truth which…wholly transcends what humans could know."[323]

5.3 Models and Approximations

Because the Absolute-PTT relies on us discovering the true theory, it is doomed. Both $PTT_{Methodology}$ ("Science should seek to formulate the perfect theory, by discovering nature's joints") and $PTT_{Quality}$ ("A theory is better the more of the perfect theory it duplicates – the more truths it contains") are untenable because the perfect theory cannot be found.

[319] E.g. Kuhn [2012], p.108.
[320] Kuhn [2012], pp.169-170 states that when *Structure* describes the scientific process, "nothing that has been or will be said makes it a process of evolution *toward* anything."
[321] Putnam [1982].
[322] He calls it "metaphysical realism".
[323] Putnam [1982], p.197.

There is, though, a response to the arguments in the previous section suggesting another version of the PTT. According to this alternative, we should not seek to find absolute truth by formulating or duplicating the perfect theory as the Absolute-PTT says; instead we should endeavour to find *approximate* truths, which can play the practical role that absolute truths cannot. I will call this the "Approximate-PTT". Psillos says that realists typically believe that "past theories are superseded by newer ones, but the successor theories are more truth-like than their predecessors".[324] Thus, theories can be approximately true, or have some "*verisimilitude*".[325] This approximation can be in different degrees, so that two approximately true theories can be compared, and thereby give an answer of question (B) of Kuhn's Challenge. Theories can also fail to have any verisimilitude at all, presumably, if their posits do not refer. This way, the basic definition of epistemological realism will need to be modified, as we cannot have perfect knowledge of the world, only approximate knowledge. But it is still possible to believe that the perfect theory is there in principle, and it is also possible to believe that our actual theories can resemble it. I will argue, however, that a notion of approximate truth that does not depend on absolute truths contradicts rather than supports the PTT.

How does approximate truths help us in comparing theories? One might be tempted to reply that that theories approximate the truth better the closer they are to the truth, but that will not do, as it brings back all the problems listed in the previous section. I will instead look at two suggested ways the notion of verisimilitude can be used to compare theories, without reference to the absolute truth.

One way to look at truth approximation is in terms of movements from the more specific to the more general, where the more general theory includes and adds the more specific, thereby achieving a better approximation of the true theory. There are certainly examples in the history of science where a more general theory has replaced a more specific. Kepler's and Galileo's laws

[324] Psillos [1999], p.280.
[325] This term is introduced in Popper [1963]. As Northcott [2013], I will use 'approximate truth' and 'verisimilitude' as synonyms.

can be seen as special cases of Newton's laws, and Newton's theory as a special case of Einstein's special theory of relativity.[326]

Unfortunately, the pattern of $\text{theory}_{\text{Later}}$ retaining $\text{theory}_{\text{Former}}$ as a special case has been the exception rather than the rule in the history of science.[327] Larry Laudan writes,

> Except on rare occasions (coming primarily from the history of mechanics), one finds neither of these concerns prominent in the literature of science. For instance, to the best of my knowledge, literally no one criticized the wave theory of light because it did not preserve the theoretical mechanisms of the earlier corpuscular theory; no one faulted Lyell's uniformitarian geology on the grounds that it dispensed with several causal processes prominent in catastrophist geology; Darwin's theory was not criticized by most geologists for its failure to retain many of the mechanisms of Lamarckian 'evolutionary theory'.[328]

The idea that earlier theories are special cases of later theories, considered as a general rule, is up against the evidence. But there is another (and more promising) way to look at relative truth approximation than in terms of scope, namely in terms of measurements. There have been several versions of this approach, but I will focus on one, the version developed by Robert Northcott.[329]

Northcott analyses closeness to the truth in terms of causes, where approximate truth, or verisimilitude, is measured as how well a scientific model captures the strength of causes present in a given situation: that is, how close a model's postulated values are to the true ones. Note that "true values" should not be understood as values postulated by a true *theory*, but as actual data measured, such as test results. The intuition is that causal strength is how much effect there is

[326] See Alex Rosenberg [2012], pp.135-141.
[327] Compare Kuhn [2012], p.168: "[N]ew paradigms seldom or never possess all the capabilities of their predecessors".
[328] Laudan [1981], p.38.
[329] In Northcott [2013].

with a given cause present, compared to the level of effect without this cause.[330] A model that has a better match to data is closer to the truth than a rival with a worse match, in respect to particular sets of causal strengths generated.

Northcott's qualification "in respect to particular sets" is important, because the measure arrived at in this way is not absolute; it is by definition context-specific. Northcott argues that this definition produces the right result for the issue of verisimilitude. He starts with the so-called "seriousness of errors" problem, where the issue is how we can compare models with a partial match to data. How serious is the deviation? Causal strength provides a useful weighting, making a model with a less significant error in terms of causal effect score better. Northcott exemplifies this with two models of a ball falling towards the Earth. Both have some shortcomings. The first model fails to account for the gravitational pull of Earth, and the second fails to account for the gravitational pull of a nearby mountain. But the causal strength of the Earth's gravitation is much higher than that of the mountain, making the second model preferable to the first, even if they both include errors.

The effect of a cause in a given situation depends not only on the cause measured, but also on the background, which, Northcott argues, corresponds to our intuition about the seriousness of errors.

> For example, striking a match may have maximum strength with respect to causing a flame if background conditions include sufficient oxygen in the atmosphere, but not otherwise. Thus the strength…is context-specific and should be understood as a token rather than type value.[331]

For Northcott, the selection of variables included in a model always is dependent on what we want to measure. The accuracy of a model is the accuracy relative to the purpose of the enquiry. Causal strength, he writes, "cannot be calculated without an exact specification of just what we

[330] The formula for causal strength is "$y_A - y_C$", where "Y" is an effect variable, "y_A" the value of Y with a certain cause present and "y_C" the value of Y without it.
[331] Northcott [2013], p.1473.

are interested in."[332] Northcott compares two ways of modelling the causes of lung cancer. The first one gives a very accurate estimate of causal strength, but mentions asbestos only, while the second includes both asbestos and smoking, but has a less accurate estimate of causal strength.

> Which of the two models should be preferred? On one hand, the first one is, as far as it goes, the more accurate of the two. On the other hand, the second one has captured more of the factors at play and so although less accurate is also more complete.[333]

There is no *absolute* accuracy as such for a model, Northcott argues, it is always a question of *relevant* accuracy. This approach agrees well with the fact that many causal models are non-linear and non-additive.[334] Therefore: "A canonical general weighting of the seriousness of errors appears impossible for one theory as a whole, let alone for science as a whole."[335]

As Northcott notes, many realists have put their hope in approximate truths underpinning their position, but his analysis does not support any sort of PTT. Approximate truth, understood in this way, is not a relative to the perfect theory, but a way to compare two models, or two theories, from a specific perspective, in relation to their empirical performance. It is not a well-defined step on the way to full truth, where the steps can be measured and compared globally: "[O]nce a theory has fallen short of full truth, thereafter there just is no univocal answer as to how *much* it has fallen short."[336] And this "full truth" we cannot have.

Northcott concludes that verisimilitude, as he understands it, is "unable to offer any particular support for the notion of global scientific progress."[337] However, it should be noted that defining verisimilitude in this way does not make it subjective or random; it gives an objective

[332] Northcott [2013], p.1481.
[333] Northcott [2013], p.1481.
[334] Northcott [2013], p.1479: "For example, when air resistance is added as a new factor to the ballistic equations, the new equation is a complicated exponential function. Therefore its distance from a postulated linear function may vary greatly, depending on a projectile's speed, wind conditions, and so forth."
[335] Northcott [2013], p.1480.
[336] Northcott [2013], p.1482.
[337] Northcott [2013], p.1487.

measurement, that is, an answer to question (B) of Kuhn's Challenge, *given a specified context*, but not a general or unique one. This will be important for me in Chapter 6.

Northcott's arguments do not make 'approximate truth' and 'verisimilitude' misnomers. There can still be something that approximate truths approximate and verisimilitude resembles. The notions are not, I will argue, incompatible with scientific realism. But as Northcott defines them, they *are* incompatible with the idea of one, perfect theory that science should aim to get closer to. This suggests a pluralistic view of science, which I will introduce in §5.6.

If this is right, approximate truths cannot be used to support the idea of a perfect theory, because either 'approximate truth' is defined in relation to the unobtainable absolute truth of the Absolute-PTT, which keeps all the problems we encountered in the previous section; or in terms of a path from the more specific to the more general, which is inconsistent with the history of science; or it implies a context-dependent pluralism that contradicts the assumption that there is a perfect theory *at all*, making the Approximate-PTT equally untenable.

There are other proposed justifications for scientific realism and I will discuss some of them in the next section. However, without the PTT, they fail to convince.

5.4 Arguments for Realism

5.4.1 Realism and Continuity

The PTT cannot serve as justification for scientific realism. But its roots go deep and it is still influential. In his book *The Disorder of Things*, John Dupré talks about (without endorsing) "the founding metaphysical assumption of Modern Western science", having in mind the positing of a "deterministic, fully law-governed, and potentially fully intelligible structure that pervades the material universe."[338] For Putnam in "MoM", the PTT is the only alternative to conventionalism and antirealism.

[338] Dupré [1995], p.2. Dupré strongly *criticises* these assumptions.

In this section I will point to how the PTT appears in some well-known arguments for scientific realism, inspired by the Kripke-Putnam world of ideas. I will evaluate them against my four criteria to see whether they can justify scientific realism, without relying on the PTT being true. But I first want to get another idea out of the way: the idea that all such arguments are superfluous because we can have realism without continuity.

Kuhn's analysis threatens conceptual continuity for natural kind terms, and this has been seen as a threat to scientific realism. Psillos, who is a realist about natural kinds, explains why realists usually are keen to defend scientific continuity:

> Why is the demonstration of referential continuity in theory change such a central element in the defence of scientific realism? Realists typically defend a cumulative approach to science…As science progresses, scientific theories offer a more refined and truer description of the world, i.e. of the natural kinds (observable and unobservable) which populate it and of their properties and causal powers.[339]

But one possible reaction to Kuhn's analysis and the issues with conceptual continuity is to say that, contrary to what Psillos thinks, there could be realism without continuity. Maybe previous theories were indeed false, and their proposed natural kind terms non-referring; but our current theories might be something else entirely and, in sharp contrast to earlier failures, capture important knowledge about reality. In respect to scientific realism, progress has been achieved by a one-off quantum leap. This is an option I am not sure that anybody has ever defended in writing, but it is available in logical space.

The option looks vulnerable to an argument that I will call "the pessimistic extrapolation argument".[340] It extrapolates from the history of false theories replacing each other, and suggests that is likely that our current theories also will be discovered to be false by future scientific activities, and that many of our current natural kind terms therefore lack reference. The same fate will eventually meet the replacement theory too, and the story will go on forever. I therefore

[339] Psillos [1999], p.280.
[340] In the literature usually called "the pessimistic meta-induction argument".

agree with Psillos that justified continuity between older and newer theories is needed to defend realism against the pessimistic extrapolation argument. I cannot criticise the realism-without-continuity option for failing to sort historical posits into referring and non-referring, criterion (**iv**), because it does so by design. Instead, the problem is that it fails to offer *any* justification for realism, since it arbitrarily holds that current sciences are radically more truth-like than older ones, and that all future scientists will agree.

Consequently, there is an issue for realism: it relies on a conceptual continuity that is problematic post-Kuhn. In the following sub-sections, I will discuss some candidates for the job to deliver the required continuity, without which we cannot justify scientific realism.

5.4.2 Two-Step Definitions

One possibility is to use an idea from *N&N* to form a response to Kuhn's Challenge and its relativistic, anti-realist implications. This idea is to take a two-step approach, where we initially use an operational (and contingent) definition and as a second step a structural (and necessary) definition, but all the time talking about the same object in the real world. This two-step idea is discussed, though not defended, by Frederick Kroon in "Theoretical Terms and the Causal View of Reference".[341]

Kripke claims that we can baptise an object with an initial operational definition, for example by using a description of its causal role, without deeper theoretical content. The role described should be read as a contingent property, although perhaps *a priori*. One example is included in a footnote to *N&N*, where Kripke describes how the reference to the planet Neptune was introduced by means of a description of its causal effects: "Neptune was hypothesized as the planet which caused such and such discrepancies in the orbits of certain other planets."[342]

The two-step idea looks promising at first, since it is plausible that the object first described by its causal effects indeed was Neptune. Maybe we can give ourselves a little wriggle room so that the characterization of objects at this initial step can be vague, relying on causal features only.

[341] In Kroon [1985].
[342] Kripke [1981], p.79, footnote 33.

This way we could stay clear of troublesome, changeable, descriptions by instead using a high-level, non-committal functional description that is doing the actual referencing. This causal-role description could then be further detailed and refined by later findings, including those of essences.

An obvious problem with the two-step idea is that it risks making reference trivially successful – a candidate natural kind term would always refer to *something*. This is not what realists like Psillos want; they want a clear difference between referring and non-referring terms in the history of science. The two-step idea does not deliver this. A functional description in chemistry such as "whatever causes combustion and calcination" would unfortunately fit both phlogiston and oxygen, and therefore give no basis for saying, as the realist wants to say, that oxygen exists but phlogiston does not.

Kripke's Neptune example is subject to the same issue, Kroon argues. That there is a lack of enough information for reference-fixing

> is evident from the way we would treat the would-be discovery that the Earth was in some very roundabout way causally responsible for these discrepancies…In such a case, we would be inclined to say that the Neptune-hypothesis was false, that there is, in fact, no Neptune.[343]

The initial operational description in terms of effects is not enough to establish a stable reference to whatever we find out is the cause of these effects. One way to provide additional support would be to say that the scientists must have been talking about Neptune all along, because the object was *in fact* eventually discovered to be Neptune. This response assumes a perspective after the event. But the issue is not whether *we today* are justified in postulating this object, but whether it was justified based on the functional definition only. For all the scientists knew at the time, the hypothesis that the cause of the discrepancies was an undiscovered planet could have been proven wrong.

[343] Kroon [1985], p.150.

Turning to ‘phlogiston’ again, a further complication is that while ‘Neptune’ is the name of a physical object, existing independently of theories, ‘phlogiston’ is a (candidate) natural kind term, naming an abstract, theoretical object. Kroon regards both ‘phlogiston’ and ‘oxygen’ as embedded in their respective theories. When we postulate phlogiston or oxygen, we explicitly or implicitly make commitments to the embedding theory. The success or failure of ‘phlogiston’ and ‘oxygen’ are connected with the fate of the theories they belong to. Say that we accept that a certain cause has a certain effect. From that we can deduce that there is something such that this something is the cause with that effect. But this falls short of what scientific realism needs. When I defined criterion (**iv**), I quoted Psillos’s idea that a scientific realism must “show that not all abandoned theoretical terms are in the same boat as ‘phlogiston’”.[344] Psillos wants to avoid the conclusion that phlogiston used to exist, but no longer does; he wants a principled way to say that some posits of old theories refer while other do not, and never did.

One way of doing this is to regard both the initial functional definition and the eventual definition in terms of essences as approximations of the *right* definition, where descriptions without such a correspondence fail to refer to proper objects – but that move would reintroduce the PTT. I will suggest a better move in Chapter 6.

Kroon takes the Neptune example from Kripke, but adds a caveat: Kroon does not claim that Kripke defends the two-step model himself. And in fact, Kripke does not. What he wrote in *N&N* instead supports the same conclusion Kroon is drawing: the dependence of operational definitions on their theoretical context. At the time of the operational definition, Kripke writes, “statements as ’if such and such perturbations are caused by a planet, they are caused by Neptune’ had the status of *a priori* truths.”[345] Kripke here argues for the success of the operational definition only given a previous classification, in this case a classification as an (undiscovered) planet. But Kroon’s point stands. There is not enough information available at the time of operational definition for a dubbing of a fully-fledged object Neptune. The classification

[344] Psillos [1999], p.292.
[345] Kripke [1981], p.79, note 33.

is tentative, waiting for a confirmation, which is not guaranteed. There is room for failures, as I will discuss further in §7.6., and there is room for decisions, which I will describe in Chapter 6.

The two-step descriptions inspired by Kripke's examples do not by themselves deliver conceptual continuity if we remove the perfect theory from the equation. The question is whether the PTT is necessary for scientific realism. I will argue that it is not.

5.4.3 The Success of Science

A second possibility for justifying conceptual continuity and scientific realism is suggested in Putnam's article "What is Mathematical Truth?", where Putnam claims that it is the only explanation for well-documented scientific success.

> The positive argument for realism is that it is the only philosophy that doesn't make the success of science a miracle. That terms in mature scientific theories typically refer…that theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories – these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science…[346]

In this quote, Putnam claims that the only explanation for the success of sciences is that:

- Terms in mature sciences typically refer; and
- Terms in mature sciences can refer to the same thing across theory changes.

There are two ways to interpret this argument. Under one interpretation, it assumes that the PTT is correct, and under the other interpretation Putnam is empirically wrong. The choice of interpretation depends on how we read the term "successful". If "successful" is to be read as "approaching the one correct description of the world", Putnam is relying on the PTT, which brings back the problems in §5.2. If we instead read "successful" to imply an evaluation

[346] Putnam [1979], p.73.

according to some empirical criteria,[347] Putnam's thesis says that historic theories that we now believe to have been more or less correct also were empirically successful, while others were not.

If we adopt the second interpretation, Putnam is up against the history of science. In "A Confutation of Convergent Realism",[348] Laudan convincingly criticises this claim using historical examples. There have been many theories, Laudan shows, where the central terms now are regarded as referential, but the theories nevertheless were unsuccessful in this sense.

> Are genuinely referential theories (i.e., theories whose central terms genuinely refer) invariably or even generally successful at the empirical level…? There is ample evidence that they are not. The chemical atomic theory in the 18th century was so remarkably unsuccessful that most chemists abandoned it in favor of a more phenomenological, elective affinity chemistry.[349]

Laudan concludes that "The realist's claim that we should expect referring theories to be empirically successful is simply false."[350] And this lack of success should not be a surprise, because

> a genuinely referring theory need not be such that all – or even most – of the specific claims it makes about the properties of those entities and their modes of interaction are true. Thus, Dalton's theory makes many claims about atoms which are false; Bohr's early theory of the electron was similarly flawed in important respects.[351]

[347] Laudan [1981], p.21, has in mind "giving detailed explanations and accurate predictions."
[348] Laudan [1981].
[349] Laudan [1981], p.24. He also mentions the Proutian theory of atoms and the Wegenarian theory about continental drift.
[350] Laudan [1981], p.24.
[351] Laudan [1981], p.24.

Referring theories were not always successful, and vice versa, many theories that were successful at the time include central terms that are now regarded as non-referring, Laudan adds.[352, 353]

As the references of the natural kind terms in those examples ('atoms' and 'electrons') are supposed to have remained unchanged, the argument for realism as an empirical theory based on the success of science is falsified. Success does not sort the historical examples in the right way for realism, as criterion (**iv**) requires. The only way to avoid this conclusion seems to be the strategy I immediately rejected, namely to commit the sin of a circular argumentation and measure success, not in empirical terms, but as progress towards the perfect theory.

5.4.4 The Principle of Charity

In "Three Kinds of Realism",[354] Putnam no longer believes that essentialism is the solution to continuity and realism, because essences are theory-dependent. Instead, the continuity on which his realism is based is now supported by the "Principle of Charity" (PoC").[355] Putnam applies the principle to the problem of scientific continuity, and in his version, the principle says that "we *should* often identify the referents of terms in different theories so as to avoid imputing too many false or unreasonable beliefs to those we are interpreting."[356]

The PoC is not as such a realist's natural tool, which Putnam realises. On the face of it, it looks as an *ad hoc* manoeuvre to make a theory come out right. But it is needed, Putnam thinks,

[352] Laudan [1981] has a long list on p.33, which includes aether theories and phlogistic theories. See also Chang [2012] for an extensive analysis of the latter theory.
[353] Laudan also discusses weaker and more sophisticated formulations of Putnam's arguments, but comes to the same, negative conclusion.
[354] Putnam [1982].
[355] Putnam sometimes calls it the "Principle of the Benefit of the Doubt". The same principle has previously been used by several philosophers in different fields. One example is Donald Davidson who uses the principle, sometimes under the name "the principle of rational accommodation", in connection with his discussion of linguistic communication. See e.g. Davidson [1973].
[356] Putnam [1982], p.200.

because "[e]quating almost any term in reference across a hundred years of growth of scientific knowledge requires the Principle of Charity in some form".[357]

In "Three Kinds of Realism", Putnam again uses a method existing already in "MoM" to motivate the use of the PoC: he designs thought experiment where we meet scientists from earlier periods. His wants to show that it is likely that we could convince them of modern ideas, and that the PoC therefore can be applied. What Putnam did to Archimedes (assigning him beliefs) he also does to Mendel. In *Meaning and the Moral Sciences*, Putnam adds: "Surely the 'gene' discussed in molecular biology is the gene (or rather 'factor') Mendel *intended* to talk about; it is certainly what he should have intended to talk about!"[358]

Putnam's conviction about what Mendel should have intended presupposes that there is a right result, which our current theories at least approximate. In this view, a justification of scientific progress *requires* the Principle of Charity. Mendel's theory must have been (a bit worse) description of the same objects as the modern theory, which is closer to the truth. But this is again the PTT view of science, which we have abandoned.

Certainly, sometimes the PoC seems to deliver the right result, for instance in the case of the electron:

> [T]here is nothing the world which *exactly* fits the Bohr-Rutherford description of an electron. But there are particles which *approximately* fit Bohr's description… The principle of benefit of the doubt dictates that we treat Bohr as referring to these particles.[359]

However, one problem with the PoC, which Putnam recognises, is that the intuitive support for its use diminishes the further back we go from the present. Connecting what Niels Bohr believed in the 1930s with today's theories is one thing, but starting with the ancient Greeks is another.

[357] Putnam [1982], p.200.
[358] Putnam [2010], p.22. (First published in 1978.)
[359] Putnam [2010], p.24.

> [E]ventually the following meta-induction becomes overwhelmingly compelling: *just as no term used in science of more than fifty* (or whatever) *years ago referred, so it will turn out that no term used now…refers*.[360]

Another major problem is that the PoC is a blunt instrument. Putnam continues: "But the benefit of the doubt can be *unreasonable*; we do not carry it so far as to say that *phlogiston* referred."[361] This means that to be able apply the PoC, we still need some criteria to decide when there is conceptual continuity, and Putnam does not give us any.

The idea that scientific progress requires conceptual continuity is sound. The idea that it requires us to see earlier terms as approximately fitting later ones is not. This latter idea only occurs to us, I suggest, if we with the PTT define scientific progress as better and better approximations of the perfect theory, which is mirroring reality. Only if we, counterfactually, have access to the perfect theory can we do our sorting and decide when the charity of the PoC should be extended, and when it should not. As the PTT is untenable, so is the PoC.

Putnam's use of the PoC was not universally well received,[362] and eventually he gave it up.

5.5 Other Types of Realism

So far I have considered justifications for a general type of scientific realism, addressing the issue with conceptual continuity – Kuhn's Challenge. I will now turn to proposed solutions that are more ontologically selective. Many writers defend versions of scientific realism that give up some of the assumptions discussed so far. I will briefly discuss some of those writers. My conclusion is that they give up too much to meet my four criteria, and I will in the next chapter go on to argue that this is not necessary. We realise this when we finally exorcise the last pieces of the PTT.

[360] Putnam [2010], p.25.
[361] Putnam [2010], p.25.
[362] Douven [1999], p.1, is particularly damning: "[the PoC] is *unjustifiable*, for as it stands it is incoherent, and even if it were not it is, barring further supplementation, unequal to its purported task."

5.5.1 Selective Realism

Selective scientific realism is restricted to a sub-set of scientific posits. It deals with both Kuhn's Challenge and the pessimistic extrapolation by giving up some realist claims, but keeping others that its advocates believe are more robust. The main types of selective realism favour entities (individuals) or structures. But from my perspective, they share a common weakness: they fail to offer a sufficient justification for realism about natural kinds.

The type of selective realism that favours *structures* is prepared to give up most ontological realist claims about the properties of entities, but asserts that there are structures in scientific theories that is not affected by theory changes. The classic formulation of this argument goes back to Bertrand Russell,[363] and identifies these structures as the formal properties of the entities, avoiding any qualitative descriptions.

> We shall say that a class α ordered by the relation R has the same structure as a class ß ordered by the relation S, if to every term in α some one term in ß corresponds, and vice versa, and if when two terms in α have the relation R, then the corresponding terms in ß have the relation S, and vice versa.[364]

Structural realism, at least in this classic formulation, is a rather spartan type of realism, as the structures only contain formal properties. The price for its spartanism, which helps it to avoid certain problems, is that we lose support for both real entities (which have properties and relations) and natural kinds.[365] Anjan Chakravartty states: "But crucially, on Russell's view, no qualitative, first-order properties or relations of the objects need be known."[366] He concludes: "By hitching its wagon exclusively to a knowledge of higher-order, formal properties, epistemic [structural realism] no longer represents a proposal for realism."[367] This conclusion is in breach of my criterion (**i**), so I will not discuss structural realism further.

[363] Russell [1927] and [1948].
[364] Russell [1948], p.271.
[365] Quine [1969] points to the incompatibility between logical sets and natural kinds.
[366] Chakravartty [2007], p.37.
[367] Chakravartty [2007], p.39.

If we instead claim that it is the *entities* that really exist, though all descriptions of these entities are theory-dependent, we again have a weapon against Kuhn's Challenge, side-stepping the problems caused by theory changes. We can believe in the individual objects without believing in the theories in which they are embedded, entity realists claim. Ian Hacking, whose entity realism goes beyond observables, describes what convinced him about the existence of electrons: they may be non-observable, but it is still possible to manipulate them. Hacking describes an experiment where a ball made of niobium is cooled and negatively charged. The charge is then changed by spraying the ball with positrons or electrons. Hacking writes: "From that day forth I've been a scientific realist. So far as I'm concerned, if you can spray them then they are real."[368]

A question is, however, whether it is possible to discover individual objects based on manipulation only, and totally without the help of theories. Chakravartty comments:

> One cannot have knowledge of the existence of entities in isolation. In order to know that something unobservable exists, one must know the details of at least some of its relations to other things – relations, for example, to instruments of detection, or to instruments of manipulation and the aspects of phenomena in which one hopes to intervene by manipulating the entities in question.[369]

He goes on to say: "In order to be a realist about entities, one must be a realist about at least some aspects of theory also."[370]

This might not seem quite right; in the Kripkean semantics we can allow the theories to identify objects using descriptions of contingent properties. This reference-fixing can work even if the descriptions in the theory are false. In the same way, we can find descriptions of relations useful to identify an object, while still not quite committing to them. But this is the difference between Kripke's reference-fixing and entity realism; for the latter, descriptions of causal properties are

[368] Hacking [1983] p.23.
[369] Chkravartty [2007], p.31.
[370] Chkravartty [2007], p.31.

not contingent, they are (in effect) descriptions of the entities' essence. Charkavartty's point is valid; what we deduce from a certain causal effect is just "there is something that causes this effect", which does not make us a lot wiser. To postulate fully-fledged objects, we need more. The point is not that postulates are fallible (this is a virtue), but that all postulates exist in a theoretical context. Without a theoretical context, entity realism suffers from the same problems as the initial step of the operational definition argument I discussed in §5.4.2.

As a result, Hacking's manipulability criterion is not conclusive; it does not remove the context-dependence. Hasok Chang points out that

> to the phlogistonists, phlogiston was not only observable (in the flame that comes out of combustion, for example), but even directly manipulable (when it was transferred from one substance to another, as in smelting or in the production of inflammable air by the solution of metals in acids).[371]

Even if we accepted Hacking's argument and extended our belief in observable concrete objects also to some unobservable concrete objects by arguing from our ability to manipulate them, or from our observations of their effects, we can obviously not directly manipulate abstract objects such as natural kinds. Furthermore, there does not seem to be an easy way to draw conclusions from individuals to kinds. To postulate a natural kind object over time we need an unchanged essence, for which we need a theoretical context with an essence type. The support for natural kinds, I conclude, is weak in entity realism too. A justification for realism that only recognises the existence of individuals and not natural kinds fails my criterion (**i**).

5.5.2 Internal Realism

A few years after "MoM", Putnam criticises what he calls "metaphysical realism", which corresponds to what I have called "the PTT". The argument is developed in *Reason, Truth and*

[371] Chang [2014], p.7.

History,[372] where Putnam calls the position of this realist: "a God's Eye point of view".[373] This criticism leads to what he there sees a better alternative, the "internalist perspective", where there is *no* God's eye point of view. For internalists, Putnam says, the question

> w*hat objects does the world consist of?* is a question that it only makes sense to ask *within* a theory or description. Many 'internalist' philosophers, though not all, hold further that there is more than one 'true' theory or description of the world. 'Truth', in an internalist view, is some sort of (idealized) rational acceptability…[374]

In "Realism and Reason",[375] Putnam says that the metaphysical claim regarding the perfect theory is *incoherent*, and that the God's eye theory cannot be distinguished from an ideal theory, defined in terms of rational acceptability. It makes no sense, Putnam submits, to talk about the possibility of the *ideal* theory still being false; that is, it makes no sense to entertain the possibility of a *perfect* theory, distinct from the ideal one.

> So let T_1 be an ideal theory, by our lights. Lifting restrictions on our actual all-too-finite powers, we can imagine T_1 to have every property *except objective truth* – which is left open – that we like. E.g. T_1 can be imagined complete, consistent, to predict correct all observation sentences…to meet whatever 'operational constraint' there are…to be 'beautiful', 'simple', 'plausible', etc. The supposition under consideration is that T_1 might be all this *and still be* (in reality) *false*.[376]

Putnam regards this position as absurd. He feels that theory T_1 *cannot* be false. To achieve the desirable goal of having a justification for realism available to us, which criterion **(ii)** requires, he thinks that we must give up another: the fallibility criterion **(iii)**.

---

[372] Putnam [1981].
[373] The problem is perhaps even worse than just our (eternal) lack of relevant knowledge. If all states of affairs for us allow multiple descriptions, we also need access to God's own language, which presumably does not allow such inelegancies (nor would of course a perfected logic, according to Frege).
[374] Putnam [1981], p.49.
[375] Included in Putnam [2010]. The text is built on Putnam's John Locke Lectures in Oxford in 1976. It should be added that Putnam in his "Author's Note" to the 2010 re-issue writes that his turn to internal realism "is one I have regarded as a mistake since 1990".
[376] Putnam [2010], p.125.

To understand the idea of an ideal theory in Putnam's sense we first need to get our heads around the assumptions that the ideal theory should predict *all* observation sentences and meet *whatever* constraints there are. If we assume that new theoretical advances and improved scientific instruments will lead to additional predictions and tests being carried out in the future, the ideal theory is forever beyond our reach. Furthermore, if we need to set aside all human limitations, T_1 requires the same God-like perspective whether we are metaphysical realists or internal realists; in both cases failing criterion (**ii**). We are just a bit worse off as internal realists, because the advantage of internal realism was supposed to be its accessibility, giving up the intuitive power that comes with the notion of a correspondence with features in nature.

If there is no point at which we have access to the ideal theory in an absolute sense, as it in practice always is open to amendments, the ideal theory needs time indices. We can perhaps make sense of an ideal theory$_{\text{Today,}}$ even if we cannot make sense of an absolute, index-free one. But if Putnam takes *that* option, the ideal theory falls well short of our realist intuitions, which identify a very important difference between what we can know and imagine today and the actual state of the world. This is why fallibility is part of my criteria for the justification of realism (**iii**). With this reading, 'internal realism' is a bit of a misnomer, because as Devitt says:

> [This theory] is not committed to correspondence truth; it is not committed to the objective existence of an external world, so far as one can tell. 'Internal realism' is not any sort of realism at all.[377]

My analysis above agrees with Devitt's conclusion. Unless an ideal theory is tied to a particular historical context and paradigm, mere humans cannot comprehend what it might be, and if it is so tied, it is unable to account for our realist intuitions.

Putnam's argument against realism only hits a realism committed to the PTT. I suggest that his alternative solution, what he calls "internal realism", *gives up* too much, landing itself in a situation where Putnam also dumps an earlier insight: the connection between realism and

[377] Devitt [1983], p.296.

scientific practices. But internal realism also *keeps* too much, as it retains a bit of the PTT: the assumption that it makes sense to talk about and aim for one fully integrated theory.[378]

Putnam is mistakenly equating (metaphysical) realism with a belief in the God's eye perspective, the assumption that comes with the PTT. But I will claim that there is no contradiction in talking about realism freed from the perfect theory.

5.6 Pluralism vs. the PTT

The PTT paints a picture of proper science according to which good theories must be compatible and well-demarcated; any deviations from this are deficiencies to be eliminated. Real progress, the PTT says, requires that sciences be integrated and consistent, and that higher-level sciences in principle be reducible to basic ones, sharing an over-arching ontology. However, this picture is far removed from actual practices also when compared to the most mature of current sciences. The current scientific landscape does not feature a homogenous, interconnected set of sciences; that much is uncontroversial. The question is whether this is a problem for scientific realism.

Along the lines of Quine's Dictum, many recent writers have suggested that it is not. We saw in the last chapter that Khalidi and Magnus both talk about scientific activities and postulates as specific to domains. Cartwright finds models in physics to be more truth-like and more stable than the explanatory theories, but no basis for an objective choice between those models; the choice is dependent on our purpose at the time. Hacking quotes from a textbook in physics, which sees no issue in stating that: "For *free* particles however, we may take either the advanced or retarded potentials, or we may put the results in a symmetrical form, without affecting the result".[379] Hacking comments: "Three models, at most one of which could (in logic) be true of the physical world, are used indifferently and interchangeably in a particular problem."[380]

If the PTT is given up, we can also give up the idea that the heterogenous nature of actual science needs fixing, and that it represents a threat or an obstacle to realism.

[378] Point made by Magnus [2012], p.114.
[379] Mott and Sneddon: "Wave Mechanics". Quoted in Hacking [1983], p.216.
[380] Hacking [1983], p.216.

Chang goes further, arguing that accepting a pluralism of multiple and incompatible theories is not only a theoretical possibility, or something that cannot be avoided; it can also be fruitful and productive for a science. To defend this view, he uses an argument from *Structure*, which says that the focus of the old and the new paradigm tend to differ; there are different questions that are relevant and deemed interesting. This can lead to a Kuhn loss, where enquiries do not get pursued after a scientific revolution. But Chang disagrees with Kuhn's preference for one paradigm at a time, and favours a pluralism where several (perhaps incompatible) projects in the same area are pursued *simultaneously*, because this leads to a situation where more issues are getting attention.[381] Specifically, Chang argues that it would have been advantageous if the phlogiston theory had continued to develop, alongside the oxygen theory. The sudden shift when the phlogiston theory was abandoned led to two disadvantages: first, some valuable problems and solution were discarded, and second, some avenues were no longer open for scientific research.

> There are two types of benefits of plurality. Benefits of toleration arise from simply allowing multiple systems simultaneously, which provides insurance against unpredictability, compensation for the limitations of each system, and multiple satisfaction of any given aim. Benefits of interaction arise from the integration of different systems for specific purposes, the co-optation of beneficial elements across systems, and the productive competition between systems.[382]

Dupré describes a pluralistic view of science where incompatible theories use their own classification systems for their own purposes, and points to different purposes existing also within a science, such as ecology and evolutionary theory in biology. His first reason, and mine, for rejecting the PTT is its incompatibility with actual, contemporary science. But the problems go deeper, Dupré states, because "most notably those [assumptions] that contribute to the picture

[381] Jones [1991], pp.198-199, disputes Kuhn's claim that the single paradigm reign is the *de facto* situation in mature sciences, and argues that it is not the case even in contemporary physics: "The diversity of ontological commitment in contemporary physics – diversity both in the nature of the things physicists claim to be committed to and in the nature of their commitments – makes any global characterization of science's activities dubious".
[382] Chang [2014], p.253.

of a profoundly orderly universe, have been shown, in large part by the results of that very science, to be untenable."[383] The three assumptions Dupré has in mind are: essentialism, reductionism and determinism. As they, according to Dupré, are untenable, there can be no unified science. I will briefly relate his argument against the first two.[384]

With "essentialism" Dupré means the claim that there are unique essences to be discovered, which I have already criticised. The reason that essentialism in this sense is untenable, he argues, is because there are always several possible classifications of reality into kinds. Dupré's example is cedar,[385] which he sees as a natural kind for a carpenter but not for a botanist.

"Reductionism" is the view that sciences form a hierarchical structure, where higher-level sciences can be reduced and explained in terms of lower-level sciences, so that the real understanding of nature is found in the description of the lowest level. Dupré is a pluralist also in respect to different levels of organisation, and therefore an opponent to reductionism. He gives a series of examples, primarily from biology, to show how implausible it is to establish the bridge principles needed for a reduction, since different objectives of sciences lead to different natural kinds.[386] In my terminology, reductionism makes sense only if we believe in the PTT, where different theories have an ideal form, and where there are natural joining points between them. But we should not believe in the PTT.

Pluralism appears well-supported. Dupré argues from the weakness of the opposition. Chang points to pluralism as a productive approach for scientific activities. Magnus expresses a scepticism against *a priori* rules for natural kinds. The proponents of such rules, disqualifying parts of science, surely have the burden of proof. But I have found no proof for that position. There are therefore good arguments in favour of concluding that the *de facto* heterogenous science is not a fault to be corrected, but stems from an inherent feature of scientific activities, which is to pursue the bests explanation within a specific field, to answer specific questions.

[383] Dupré [1995], p.2.
[384] For completeness: Dupré sees determinism as undermined by the logical obstacles to prediction caused by chaos theory.
[385] Compare my discussion about jade in §4.7.
[386] Compare my discussion about planets in §4.7.

Supporting pluralism of course does not imply proposing an *a priori* rule *against* reductions in science – there are many successful examples of reductions. But it does leave it up to the scientific communities to find the optimal tools for their respective explanatory tasks.

5.7 Scientific Realism Nevertheless

In the previous section I found arguments in favour of pluralism, in opposition to the assumption of a homogenous science that comes with the PTT and leads to $PTT_{Methodology}$.[387] Does this mean that we need to give up scientific realism? Both in "MoM" and in later writings Putnam believes that we have to choose between the PTT and antirealism. Some writers further argue that the unity of science is needed for realism. But others maintain that we can accept pluralism and still be realists.

Neither Kuhn nor the later (antirealist) Putnam excludes the existence of an objective reality altogether. In particular, their criticism of the PTT is compatible with the existence of an objective reality with which we interact, and about which we are able to learn from this interaction. As this takes place, for instance in the testing of scientific theories, there is an influence of the real world on the sciences. This is the case also for Northcott's version of verisimilitude, which holds models to be approximately true in relation to the hard facts of test results. If a critical test fails, this is naturally interpreted as feedback that the combination of theories subjected to the tests (assuming that instruments function as expected) is not a good description of the world.[388]

However, feedback from the world does not establish that any particular theory is correct.[389] The world underdetermines the conceptualisation and the decision-making. But that is only really serious for the PTT. There are other sorts of realism. A non-PTT type of realism, with an emphasis on the fallibility of our theories, has no problems with pluralism. Indeed, I claim that a

[387] "Science should seek to formulate the perfect theory, by discovering nature's joints."
[388] Chang [2014], p.215, talks about "nature's *resistance*"; Dupré [1995], p.13, about theoretical posits that can only be understood as "interacting with a real and sometimes recalcitrant world."
[389] We would know that *something* is wrong – but not that a specific hypothesis is wrong, due to the dependence on other factors. This is often called "the Duhem thesis" after the physicist Pierre Duhem.

realist who has rejected the PTT should embrace pluralism and the endeavour to capture a likeness of reality in different ways: to answer a variety of questions about the world from many perspectives.

This view gains support from the pluralist writers I discussed in the last section, who defend versions of realism that are far removed from the PTT. For example, Dupré offers a pluralistic, "promiscuous" realism that allows that several different theories all can be (better or worse) descriptions of the same world.[390] What Dupré says about the subject matter of biology applies to the rest of creation too:

> There is no God-given, unique way to classify the innumerable and diverse products of the evolutionary process…Realism about biological kinds has nothing to do with insisting that there should be some unitary cause of biological distinctions.[391]

For Dupré, the pluralism of incompatible scientific descriptions is anchored in features of the world, as he recognises many distinctions between things. Similarly, Jonathan Cohen and Craig Callender describe their position in this way:

> [T]he world permits possibly infinitely many distinct carvings up into kinds, each equally good from the perspective of nature itself, but differentially congenial and significant to us given the kinds of creatures we are, perceptual apparatus we have, and (potentially variable) matters we care about.[392]

Hacking also puts pluralism on a metaphysical footing when he pictures the Creator as the author of "a Borgesian library, each book of which is as brief as possible, yet each book of which is inconsistent with every other. No book is redundant."[393]

These authors hold that pluralism does not rule out scientific realism. Chang states that for his version of realism, "active realism", pluralism should be strongly encouraged and that a number

[390] Dupre [1995], p.17: "Definitions are easy enough to draw. Useful ones are another matter."
[391] Dupré [1995], p.57.
[392] Cohen and Callender [2009], p.22.
[393] Hacking [1983], p.219.

of systems of practice should be cultivated "as incommensurable as possible from each other!"[394] Pluralism recognises multiple theories and data as (fallible) knowledge of the real world, even if they are not combinable, and as a research methodology it supports an increase of this knowledge.

For Chang, realism follows from his pluralism: "[R]ealism should be taken as a commitment to maximise our learning from reality [i.e.] whatever is not subject to one's will".[395] And: "The active realist ideal is not truth or certainty, but a continual and pluralistic pursuit of knowledge."[396] In the same spirit, Peter Godfrey-Smith formulates a version of scientific realism that gives an alternative to the PTT view, a version compatible with pluralism. He stresses the aims of science in his definition of scientific realism, and separates this definition from the tricky issue of which scientific posits we should believe in, and commit to, in our ontology. On top of a metaphysical realist assumption, Godfrey-Smith adds:

> One actual and reasonable aim of science is to give accurate descriptions (and other representations) of what reality is like. The project includes giving us accurate representations of aspects of reality that are unobservable.[397]

Godfrey-Smith's definition implies that scientific realism should conform to scientific practices ("actual") and that scientists can hope to be successful from time to time ("reasonable"). And importantly: the definition does not rule out that science can provide more than one explanation; it does not require a unique description. In addition, Godfrey-Smith disconnects the issue of a commitment to realism from the actual success of scientific theories. His definition says "accurate" rather than "true", which at least partly is due to the fact that Godfrey-Smith recognises also non-linguistic representations of the world. He does not rule out that sciences can have more aims than to give accurate descriptions of reality.

[394] Chang [2014], pp.217-218.
[395] Chang [2014], p.203.
[396] Chang [2014], p.203.
[397] Godfrey-Smith [2003], p.176.

To what extent we should commit the postulates of sciences to our ontology is for Godfrey-Smith a much more nuanced and complicated matter than a commitment to realism, since the entities or the structures postulated by sciences can be strong candidates in different contexts. "We might find good reason to have different level of confidence, and also different kinds of confidence, in different domains of science."[398] If certain models in physics are worthy of ontological commitment when we do physics, this has no obvious implication for biology or sociology. The claim to commitment can vary according to our scientific projects; according to what is vital for our best answers to particular questions of interest. Being a scientific realist in this sense is respecting our realist intuitions but leaving the ontological postulating to different scientific enquiries, for their specific purposes.

There are close connections between pluralism, implied by Quine's Dictum, and realism. They are not only compatible when the PTT is removed; realism without the PTT naturally leads to pluralism, since we should try to understand reality in as many ways as we can.

5.8 Conclusion

One type of scientific realism, the PTT, claims that there is, in principle, a perfect theory which is true, as all its postulates correspond with features in the world. The PTT promises to help define scientific progress: it consists of a better and better approximations to the perfect theory. But if it is based on absolute truth, the PTT cannot serve as a justification for scientific realism because the perfect theory is forever unavailable. Turning to approximate truths, we found that they cannot do the extra-theoretical job the PTT expects of them.

I discussed three types of arguments, defended or inspired by Kripke and Putnam, used in favour of realism. But I found that none was powerful enough on its own[399] to justify scientific realism according to my four criteria, unless implicitly relying on the PTT. Also the types of selective realism I discussed next fail at least one criterion.

[398] Godfrey-Smith [2003], p.178.
[399] I will cash in this proviso in the next chapter.

Looking for a response to Kuhn's Challenge and a justification for scientific realism, I chose a methodological principle that I called "Quine's Dictum". The Dictum might at first glance appear to be incompatible with scientific realism. In particular two issues look problematic: the heterogenous state of science and the role of decision-making. I discussed the first of these issues in this chapter and concluded that pluralism is incompatible with the PTT; a realism without the PTT, on the other hand, instead implies pluralism, an approach that I found well-supported. It is possible to both respect for the heterogenous nature of current sciences demanded by Quine's Dictum, *and* be a fully-fledged realist, as soon as the PTT is given up.

I discuss decision-making in the next chapter.

Having gone through most of the chapter with only an intuitive understanding of what scientific realism is, when the PTT is rejected, I at the end agreed with Godfrey-Smith's definition, which separates the issue of realism from the issue of ontology. Godfrey Smith gives a good working definition of scientific realism, which is also compatible with Quine's Dictum. There is however something important missing from the definition, namely the Kripke-Putnam insight about natural kind terms: they are rigid and scientific identifications are necessarily true (not just "accurate") if they are true at all. More generally, what is missing is an account of the conceptual continuity on which scientific realism depends. I will return to this in the next chapter.

6 Change and Continuity

6.1 Introduction

I have discussed realism and argued that perfect-theory realism is not the only type, and that realism can be compatible with pluralism. I also argued that realism is dependent on conceptual continuity. In this chapter, I therefore return to the issue I started with: how to respond to Kuhn's Challenge.

The PTT implies that there can be general rules across sciences for what constitutes scientific continuity; it is achieved when a theory keeps elements of an earlier theory that approximated the true theory. We make progress when this proximity is further advanced. In this chapter, I follow Kuhn in questioning whether there are any such general criteria at all. For him, the choice of the scientific community in question is the last word, as "there is no standard higher than the assent of the relevant community."[400] But I add that this is no threat to scientific continuity and therefore no threat to scientific realism.

Central to this discussion is how scientific activities bring about the changes that Kuhn describes. In particular: how are essence types chosen if, as I have argued, there is more than one legitimate essence type used for natural kinds? According to the PTT, the essences of natural kinds are *discovered*, as scientists unmask the secrets of nature. But in his book *Natural Kinds and Conceptual Change*, Joseph LaPorte, in Kuhn's spirit, argues that decisions rather than discoveries have the key role. I will introduce LaPorte's ideas with a handful of examples. My conclusion is that if we continue to follow Quine's Dictum, our account of scientific realism also needs to cover decision-making. But the lack of general rules, available *ex ante*, does not make the decisions arbitrary. And the role of decision-making does not rule out that scientific progress involves discoveries.

[400] Kuhn [2012], p.94.

At the end of the chapter, I will return to one of Putnam's two mechanisms for conceptual continuity that I discussed in Chapter 3, the historical-chain argument, and offer the underpinning it needs to be tenable. The continuity of reference needed for scientific realism can be defended, I argue, if we recognise the role of decisions taken for good reasons.

6.2 The Role of Decision-Making

LaPorte writes: "I argue that scientists' conclusions are not, in general, discovered to be true…They are stipulated to be true. Contrary to the received view, scientists change the meanings of kind terms."[401] He backs this up with a series of examples from the history of science. I will consider some of them.

As the etymology of 'ruby' suggests, this stone was originally believed to be red.[402] Chemically, ruby is a crystalline form of aluminium oxide (Al_2O_3) called "corundum". It has been regarded as a valuable gem since antiquity.[403] When this underlying microstructure of the gem was eventually identified, it was discovered that the same mineral also was the microstructure of gems with a range of other colours.[404] The name 'ruby' was not extended to the other colours, but kept for the red variety only. LaPorte concludes that it would not

> be right to say that "ruby" was *discovered* to be a variety of the mineral that composes it, corundum. On the contrary, whether "ruby" can be blue was unclear when speakers first learned that there are other colors of the relevant mineral.[405]

The story about 'topaz' is very similar to the story of 'ruby', but with the opposite outcome. Originally, all topaz stones were believed to be yellow, but when it was found that the underlying microstructure also came in blue, the blue stones were recognised as real topaz. LaPorte finds his

[401] LaPorte [2004], p.2.
[402] Latin 'ruber', meaning 'red'.
[403] Albeit not as valuable as wisdom or a good wife, according to *Proverbs*.
[404] It is called "padparadscha" if pink-orange, and "sapphire" if it has any other colour. To be pedantic, *all* ruby colours emanate from impurities.
[405] LaPorte [2004], p.102.

thesis supported: "That 'topaz' refers to all of one mineral and 'ruby' to only the red of another seems to represent decisions, not discovery."[406]

In "MoM", Putnam claims that the history of 'jade' is importantly different from the story about 'water' in his Twin Earth thought experiment. But LaPorte argues that Putnam is empirically wrong, and that the true story brings the 'jade' case much closer to the 'water' case – although with another outcome, based on a conscious decision. Putnam writes that 'jade' for a long time was used for two different minerals.[407] But in fact, LaPorte says, Chinese speakers had for many years been familiar with jade, by which they meant gems from the mineral nephrite, when they at one point

> encountered a new substance with properties similar to those of a formerly recognized substance but with a completely different microstructure, and speakers responded by applying a term for the formerly recognized substance to the new substance.[408]

Although craftsmen who were used to working on nephrite jade could tell the difference between nephrite and the new type of mineral, called "jadeite", the result of the discussion was not to give up using 'jade', nor to restrict it to nephrite only,[409] but to include both. Gems made from both nephrite and jadeite were (and still are) regarded as true jade. LaPorte adds: "Amazingly, jade met its XYZ."[410]

The point of this example is not, I believe, that jade has a disjunctive microstructure as its essence (jade = (jadeite or nephrite)), nor that jade is not a proper natural kind. There is another alternative: we can instead utilise the possibility of multiple essence types I outlined in Chapter 4. We can say that the vernacular natural kind jade has an essence consisting of functional

[406] LaPorte [2004], p.102.
[407] LaPorte is actually not fair to Putnam. Putnam's point is not related to the order in which the minerals were discovered but to the actual occurrence on Earth where the reference of our use of 'water' and 'jade' was fixed. See Putnam [1975], p.241. I ignore that to focus on LaPorte's view on decision-making.
[408] LaPorte [2004], p.95.
[409] LaPorte points out that the elimination of a natural kind term from the taxonomy in the light of new findings is one of the roads open to science (e.g. 'reptiles' and 'caloric'). This did not happen to 'jade' when the difference in microstructure was discovered – but it could have happened.
[410] LaPorte [2004], p.95.

properties and that its (disjunctive) microstructural property is contingent. The advantage of that approach is that we get an explanation of the decision to keep the name for both minerals, which otherwise would be a mystery.

The situation for 'water' in on Earth 1750 was very similar to that for 'jade' before jadeite was discovered, LaPorte argues. At that time, he claims, 'water' was *vague* in respect to XYZ; had XYZ been discovered, it would have neither been clearly inside or clearly outside the extension of '$water_{1750}$'. The vagueness would be ironed out after a discovery, as also happened with 'ruby', 'topaz' and 'jade', but that required a refinement of what 'water' means: "Contrary to Kripke and Putnam, 'water' would change its meaning."[411,412] Such a refinement of the meaning would take a decision, as it did in the 'ruby', 'topaz' and 'jade' cases.

> If we were to find XYZ, we would have found a substance that is a borderline member of the extension of 'water'. We might call XYZ "water", contrary to Putnam. Then again, we might not: We could go either way.[413]

In a modified version of the Twin Earth thought experiment, LaPorte lets the Twin Earth liquid have D_2O (deuterium oxide) as its dominating microstructure.[414] His example deviates from Putnam because the observational properties of H_2O and D_2O differ slightly at a close look, and importantly, D_2O is a close relative of H_2O.[415] Our intuitions might now be different than in the original version, and LaPorte suggests that the Twin Earth substance could equally well have been regarded as a separate substance or as another type of water; there was no right or wrong.

These conclusions seem inconsistent with conceptual continuity, and therefore with scientific realism. But I will now argue that "could go either way" is not a happy phrase to describe

[411] LaPorte [2004], p.93.
[412] In §4.4. I quoted Zemach's observation that the extension of 'water' in the past has included many non-H_2O liquids.
[413] LaPorte [2004], p.100.
[414] Deuterium oxide is more commonly called "heavy water", and it *is* a form of water. It occurs naturally in small amounts in Earth water.
[415] LaPorte's example appears more realistic, avoiding the Kuhn/Dupré criticism of Putnam's example (see §3.5). But that is an illusion, I will argue in §7.8.

scientific decision-making. Such decision-making, better understood, instead of threatening conceptual continuity gives us a tool to *defend* it.

In the next section, I will continue to discuss the role of decision-making in science based on a number of case studies, but add a claim: these decisions were *taken for good reasons*. Recognising the role of good reasons, specific to each case, helps us to address Kuhn's Challenge.

6.3 Decisions for Good Reasons

6.3.1 The Minerals

LaPorte convincingly argues for the important role of decision-making in theory changes. But in his comments, he sometimes goes too far. His analysis of the ruby case continues: "Speakers could *as easily* have started applying 'ruby' to the other colors of that mineral *as not*."[416] LaPorte's choice of words here suggests chance and randomness, and he has therefore been criticised for implying that there were no bases for these decisions.[417]

This is how LaPorte describes the background to the ruby decision:

> When it was eventually realized that the things called "ruby", all of them red, were specimens of a mineral that comes in many colors, people nonetheless continued to reserve 'ruby' for the red specimens of that mineral.
> Speakers were able to continue to call only red stones "rubies" not by ignoring science but rather by interpreting 'ruby' as a name for a mineral *variety* instead of an entire species.[418]

One might feel that the process is underdescribed; *why* did the speakers chose to interpret the scientific discoveries in this way? The same question arises for LaPorte's description of the corresponding topaz process: "In this case speakers responded differently than in the ruby case.

[416] LaPorte [2004], p.102. Italics added.
[417] See Bird [2010].
[418] LaPorte [2004], pp.101-102.

Speakers concluded that blue specimens of the mineral *are* topaz."[419] One suspects that the explanation of the difference instead is the presence of decisive factors other than the microstructure. Similarly, the conclusion from the jade example is surely not that the decision to keep the common name for both minerals was due to chance, but that the factors used to determine the outcome in this case were specific (including, I assume, factors such as aesthetic and commercial value).

If this type of explanation is along the right lines, it has an important implication. If the illuminating type of explanation is specific to one decision, it is not necessarily possible to generalise. The reasons that led to 'jade' being used for two types of minerals cannot necessarily explain any other decisions, although they were rational and decisive for this one.

LaPorte quotes Geoffrey Wills's book *Jade*, but without elaborating. Wills writes: "In the main, it is correct to say that the differences between jadeite and nephrite are of greater interests to archaeologists, mineralogists and geologists than they are to collectors."[420] That is exactly the point. It is the relative weighing of the interests of different stakeholders in this particular natural kind term, which determines the outcome. We do not need to say that it "could have gone either way" (which LaPorte says), and even less that the outcome was due to chance (which he does not say). A deity with perfect knowledge could perhaps have been able to predict the eventual decision for 'jade', having full knowledge of relevant factors and their weighting, but He could not have found that this formed a common pattern applicable also to 'water', 'ruby' and 'topaz'. A good reason in one case can be an inadequate reason (or entirely irrelevant) in another case. Building on the conclusions from Chapter 5, 'jade', 'ruby' and 'topaz' are all natural kind terms, but for enquiries with different purposes. Topaz and ruby are natural kinds in chemistry – but jade is not. Classification as jade does however come with explanatory power, supporting generalisations in another context, the context of jewellery, which is reflected in the vernacular term. The idea of one *unique* context for *proper* natural kinds should be given up with the PTT.

[419] LaPorte [2004], p.102.
[420] Quoted in LaPorte [2004], p.99.

6.3.2 Water

I will now discuss water (and 'water') from another angle, in a story about decision-making told by Hasok Chang. Staying on Earth and sticking to actual history, Chang describes the way science arrived at the identification 'water is H_2O'. He describes an iteration between theory and data that is far from a straightforward discovery, even when the relevant elements had been identified. With modern chemistry and the arrival of chemical atomic theory, water came to be recognized as a compound consisting of oxygen and hydrogen, and no longer as an element as the phlogiston-based theory had insisted. But in which proportion of oxygen and hydrogen?

John Dalton had early in the 19th century no easy way of working out the proportions without knowledge of the relative atomic weights (1:16), because he could only directly observe gross combining weights. Had he known the molecular formula, he could have inferred the atomic weight from the combining weights. Alternatively, if he had known the atomic weight, he could have inferred the molecular formula. But he knew neither and had no means to find out, so he was stuck. Chang adds: "We can make up any self-consistent system of atomic weights and molecular formulas, and observation cannot refute our system."[421] Dalton needed a principle to break this deadlock, and he went for simplicity and a one-to-one relation between the two elements, that is HO. He motivated that by the belief that other proportions would make the molecule less stable. This solution did not satisfy others, who realised that Dalton's motivation was unsound and that his method did not work for molecules with more than two elements. It was also realised that there were *other* compounds of oxygen and hydrogen (such as peroxide) and that both elements were part of many other compounds. A solution for atomic weights and molecular formulas could therefore not be found for water in isolation.

This was a problem in chemistry for fifty more years. Competing theories were put forward during the first part of the 19th century, until a solution was found via an auxiliary hypothesis stipulating a relation between atomic weight and the better understood atomic volume: in this case a 2:1 volume ratio between hydrogen and oxygen in water.

[421] Chang [2014], p.139.

Chang stresses that there were many possible ways to go; the theory was underdetermined by evidence. Also a system based on water being an element rather than a compound, Chang argues, could be cogent, as

> there are perfectly rational and sane conceptual universes, fully informed by modern science, in which water is an element, or it is a compound of some other constitution than H_2O. And these 'conceptual universes' are simply different ways of thinking about and dealing with the actual universe we live in.[422]

Like LaPorte, Chang recognises the key role of decision-making, and the possibility of different outcomes. But as I will argue in the next sections, this possibility does not make the actual outcome arbitrary.

6.3.3 Phlogiston vs. Oxygen

I have discussed Chang's analysis of the role of water in the dispute between the phlogiston-based and oxygen-based theories, and finished with his conclusion that there were alternatives, more than one way to go. Chang argues further that during the Chemical Revolution, the oxygen theory had not by itself any decisive advantage over the phlogiston theory. But the decision to change was still taken for a reason; they did not throw dice. According to Chang, the choice between alternative theories depended not just on the theories themselves, but also on general directions of scientific thinking.[423] An example is the transition from a tradition Chang calls "principlism", with an explanatory model based on general principles, to "compositionism", which posits stable components rearranged in chemical reactions.[424] The latter was more congenial to Lavoiser's oxygenist theory since it gives a key role to weights.

[422] Chang [2014], p.209.
[423] I will discuss this further in §7.4.
[424] Chang [2014], p.37, states that principlism underpins phlogistonist doctrines, because such theories "incorporated a significant metaphysical doctrine about the fundamental ontology of chemical substances, which differed from each other sharply." In contrast, (p.38) the important principle for compositionist theories such as Lavoisier's, was to describe chemical substances as "either elements, or compounds made up of those elements."

> [T]he clear evidential advantage of the oxygenist system on the basis of weight considerations only holds if one accepts compositionism; phlogistonists disregarded weight-based arguments because they were principlists. The Chemical Revolution makes much more sense when we see it as a ripple riding on a large wave, which was the very gradual establishment of compositionism.[425]

The fact that oxygen and phlogiston were not just key posits in competing theories, but belonged to different ways of thinking about chemistry, following different principles, gives the background to the discontinuity. Phlogiston is not oxygen under an updated description; according to our best knowledge, there is no such thing.[426]

In this situation, there were more than one way to go for chemistry, as Chang shows, but this is not the whole story. There was a good reason to choose H_2O, although this reason was not just due to empirical data. Once the key role of weight was accepted, as part of accepting compositionism, the conclusion for water followed, and this determined the way the scientific society was going. An identification with HO does not fit with that system.

> As long as we are working within certain systems, not believing H_2O is of course going to create some incoherence in our system of practice. It would not have worked to practice organic structural chemistry after the 1860s while maintaining HO.[427]

An example from biology supports the same conclusion.

6.3.4 Species[428]

In Darwin's days, the reigning orthodoxy, the special creationist theory, held that:

1. The division into species should be close to what was normally regarded as such: Bears, dogs and radishes are examples of species.

[425] Chang [2014], p.42.
[426] Chang [2014], p.45, mentions another possible connection: "[I]n a whiggish understanding of the phlogistonist theory of metals, there is a clear reason to identify phlogiston with free electrons."
[427] Chang [2014], p.187.
[428] In this section I follow LaPorte's [2004] exposition.

2. The species were created as they are now; they have not changed significantly.
3. A species includes all blood relatives of an original population.
4. The species were all created at the same time.

Before Darwin, the natural kind term ‘species’, as it was used by the special creationists, included the assumptions used by that theory, to the extent that some regarded it as a contradiction in terms to say that species have arisen by evolution. The accepted definition included elements (1)-(4) above.

It had however become an established fact at that time that the fourth thesis was incorrect: new plants and animals had been introduced over time while others had become extinct. The group of assumptions had to be adjusted.

Initial options were:

- Continuous creation (assumption 4 is given up).
- Evolution from more primitive animals and plants (assumptions 2-4 are given up).

Ruling out the first option, which kept more assumptions but required regular miracles, there was now agreement on the facts. That left the question about terminology, about how ‘species’ should be defined.

Terminology options were:

- Retire ‘species’ as being vague and therefore not useful.
- Restrict ‘species’ according to assumptions 2 and 3 and say that it does not refer.
- Give up assumption 1 and allow only *one* species, including all living things.
- Restrict ‘species’ according to assumption 1 and give up assumptions 2-4. This was Darwin's proposal, which was accepted. After this meaning change, ‘species’ is scientifically useful in the theory of evolution.

The continuity of ‘species’ was questioned when it underwent this radical change. LaPorte quotes Darwin’s contemporary critic William Hopkins, who insisted: “Every natural species

must by definition have had a separate and independent origin". Hopkins therefore drew the conclusion that Darwinists "in fact, deny the existence of natural species at all."[429]

Some fundamental assumptions were changed, and a new, historical essence type was introduced in phylogenetic theories. But there was also continuity; the new concept referred to the same type of entities as the old because assumption 1 was kept. Species, we would like to say, have always existed, but were previously *wrongly described*; we choose continuity, with an updated definition. The meaning of 'species', LaPorte claims, has had an increased precision in the relevant respects. Statements like 'New species have arisen by evolution' that were vague in the old theory, given the new data (as multiple options existed) are true in the new theory. But this took a decision, as LaPorte argues. The change of meaning was not the only option consistent with the data, but one that made it useful for the new theory.

In my earlier discussion about such theories, I already quoted Dupré's criticism of extra-theoretical essentialism in relation to species: "Phylogeny, in short, cannot possible create essences *ex nihilo*."[430] Starting points have to be chosen. This supports LaPorte's point: there is still a key role for decision-making when defining species. Yet the decisions are not arbitrary.

'Species' is a complicated case, as there are still many competing definitions. Darwin wished to keep the division into species close to the vernacular classifications, and the same wish led to the phylogenetic theories. Structural definitions, both morphological and genetic, struggle in this respect. On top of this, all definitions want to offer scientific advantages for their purposes of enquiry.

Furthermore, LaPorte points to the need for decision-making also for classification of individual species. The guinea pig is no longer counted as a rodent. A guinea pig looks rather like a hamster and was therefore earlier grouped with hamsters, mice and rats. Empirical data do not dictate that this cannot be done, LaPorte writes, but set constraints for the choice. In this case, any group that includes guinea pigs, hamsters, mice and rats *also* needs to include many other animals, for

[429] LaPorte [2004], p.124.
[430] Dupré [1995], p.57.

example horses, seals and primates, once we accept a phylogenetic approach. If we start a species with the latest common ancestor to hamsters and guinea pigs, this ancestor would also have *homo sapiens* as a direct descendent, making us all part of that same species. This was in principle an option open for biological taxonomists, and sometimes the decisions in biology have gone for extensions rather than reductions. But here, it is clearly the worse option. There are obvious practical reasons against a classification that makes us all rodents.

6.3.5 Planets

I earlier referred to Magnus's discussion of planets as natural kinds for astronomy. The natural kind term 'planet' and the case of Pluto also illustrate the disciplinary constraints for decision-making.[431]

The International Astronomical Union (IAU) adopted a new definition of 'planet' at a conference in 2006. Magnus quotes from IAU 2006:

> 'A planet is a celestial body that (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighbourhood around its orbit'.[432]

One of the effects of this resolution is that Pluto is no longer classified as a proper planet (it is now a "dwarf planet"). This was not an arbitrary change of terminology; it was not just about the use of a word. The background to the decision was that new findings had established that Pluto is one of many similar, trans-Neptunian so-called "Kuiper Belt objects". In 2003, one such object significantly *larger* than Pluto was discovered. The situation needed to be addressed. The choice was between having eight planets, excluding Pluto, or many more. We again had a situation where the traditional use of a term no longer made sense given new, important data, as was the case for 'species' in Darwin's days.[433] Also for 'planet' there was a decision needed, but it was a

[431] This example is described in Magnus [2012], and I follow his exposition.
[432] Magnus [2012], pp.73-74.
[433] A "scientific crises", in Kuhn's terminology.

decision with constraints: "The crucial point is that no scientifically viable choice would have preserved the familiar list of nine planets."[434]

The choice between eight and many was not obvious, but it was not arbitrary either. It followed from the conditions in the IAU definition, chosen from the perspective of what was useful for the astronomy of solar systems. In particular, the condition that requires that a planet must have cleared its neighbourhood, meaning that it must be the dominating gravitational body around its orbit, was crucial. Magnus continues:

> [A]ll the conditions in the IAU definition of 'planet' pick out features that are astronomically significant…[A]adopting a weaker definition of 'planet' [allowing more objects to meet the conditions] would have come at a cost in explanatory success.[435]

The result was a clarification of a term that was found vague in the light of new discoveries, amending the meaning of the term (what Kuhn calls a "redubbing"[436]), but keeping the continuity in a way that makes sense for the study of solar systems. There was a reason for the change.

6.4 Discoveries vs. Stipulations

As we have seen, LaPorte argues that theoretical identifications typically are the result of decisions rather than discoveries, and decisions where more than one outcome was possible. This is not so far from Kuhn's view, according to which there is no real difference between inventions and discoveries. Kripke and Putnam, on the other hand, talk about scientific breakthroughs as discoveries. Do the above cases show that we must reject a role for discovery?

No. The distinction between essences and essence types makes it possible to agree with all of these claims.

[434] Magnus [2012], p.83.
[435] Magnus [2012], p.78.
[436] Kuhn [1990].

The theory-dependence of essence *types* is clear. That observable objects have a microstructural essence was not discovered in any straightforward way, and the same is true for Cladism, which stipulates essences of a historical type; both were hypothesised by a theory, a theory chosen for its actual and potential explanatory power. In Chapter 4, I argued against a general primacy of microstructural essences for scientific purposes, and pointed out that also functional, macrostructural and historical essences are used to good effect.

That the essence of water is H_2O, on the other hand, was a discovery to be made, as was the ancestry of tigers, *given the choice of essence type*. A neurological analysis would just be looking in the wrong place for mental states if their essence type is functional, while a functional analysis could generate discoveries. A search for planets among celestial bodies is not well served by an analysis of their chemicals – but a search that applies the IAU definition is. No amount of normal-science research will by itself change that in the future, although empirical discoveries can lead to reconsiderations of definitions and to redubbing events.

I earlier referred to the non-trivial choice of essence type for biological species. This biological example also shows something surprising. An essence is understood to be a property that an object must have to be that object. But if considered as a necessary condition for continuity across paradigm shifts, even that reasonable-sounding candidate is too strong. Historical examples suggest that a property might be regarded as essential by one theory, but contingent by another, because the essence type has changed. Still, the outcome might be to continue to use the same natural kind term, with another essence, as was the case for ‘species’.

Putnam makes a very similar point in “Three Kinds of Realism”, where he discusses the relation between Newton’s and Einstein’s theories in terms of their core properties, and puts forward a Kuhnian argument:

> Special relativity preserves many notions from Newtonian physics (while making them frame-relative): e.g., *momentum*, *kinetic energy*, *force*. We can view special relativity as preserving the ‘core’ of Newtonian physics *if* we take the ‘core’ to be the approximate correctness of the Newtonian laws at ‘non-relativistic’ distances and speeds (i.e., speeds

> small in comparison with light, and distances small in comparison with a light-second). But this would be a totally arbitrary way to define the 'core' of Newtonian physics from a *Newtonian* point of view.[437]

Essences, or core properties, for scientific natural kinds are theory-dependent. Using them across theories does not make the story extra-theoretical, only ahistorical. Neither contingent nor essential properties *necessarily* survive theory changes, I conclude. Decisions, made for good reasons, trump essences.

6.5 Historical-Chain Continuity Ex Post

The above conclusion might seem to threaten the conceptual continuity required to justify scientific progress once again. However, I propose that by deploying the idea of decision-making for good reasons, we can develop a more plausible version of Putnam's historical-chain argument for conceptual continuity. In Chapter 3, I described this argument as saying that "we can bypass all (other) meaning components by relying on an unbroken chain of reference following an initial baptism."[438] As we saw, the problem with the historical-chain argument on its own is that it relies on the $same_x$ relation, which in turn relies on an implausible view of unchanged, trivially identified essences, and on an equally implausible view of meaning-constancy over time. But if we accept the crucial role of decision-making for good reasons, those views are not required. The continuity is provided by the good reasons of the decision-making in the scientific communities. These reasons are not general and *ex ante*, however, they are specific and *ex post*. I will show this by again pointing to examples from the history of science that I have already discussed, before introducing a longer and more complex story.

In the middle of the 19th century, a decision was needed for *species*, as new discoveries meant that the special creationist theory was not consistent with relevant data, and some of the assumptions in the older theory had to be given up. The choice of a new theory was constrained by the need to adjust the theory so that this inconsistency was removed. There was a choice of

[437] Putnam [1982], p.199.
[438] §3.2.2.

which assumption to give up, and more than one alternative existed. *Continuous creation* would also have explained the data available, but this would sit badly with the general principles of post-Darwinian biology. A *single species* for all living creatures would also have removed the problems, but be drastically at odds with the vernacular classification. Another alternative would have been to give up species as a natural kind, but in the end, the natural kind term 'species' was adjusted in such a way that it offered explanatory power also in Darwin's theory.

For *water* in the 18th century, the phlogiston-based system became increasingly difficult to reconcile with empirical results and a new decision was called for. More than one option could have addressed this issue. But the fact that there were many possible theories compatible with the available data did not make the theories equally good. If the context of other, related theories and more general movements and practices in chemistry are taken into account, the eventual decision is well justified. Chang's analysis of the pros and cons of phlogiston-based and oxygen-based theories does not result in a decisive verdict on which theory has the better fit to data. But the choice was still made for a good reason; it followed the change of a guiding principle in chemistry at the time, from principlism to compositionism.

For *planet*s, the status for Pluto became difficult to defend when new bodies, larger than Pluto, where found in our solar system. Again, a decision was needed, and again, there was more than one option. A definition was agreed based on what astronomers needed, resulting in a choice of the practical solution, the demotion of Pluto.

The decisions I have described did not follow general *ex ante* rules. Yet they suggest there is real, justifiable extra-theoretical continuity, not in the form of general *ex ante* rules, but in the form of specific *ex post* stories that accommodate the important role of decisions as well as discoveries. The continuity defended here is not just the continuity in Wittgenstein's rope,[439] but continuity within the framework of knowledge and methodologies for each individual purpose of

[439] Wittgenstein [1958], p.87: "[when we have understood the issue of similarities] we no longer feel compelled to say that there must be some one feature common to them all. What ties the ship to the wharf is a rope, and the rope consists of fibres, but it does not get its strength from any fibre which runs through it from one end to the other, but from the fact that there is a vast number of fibres overlapping."

enquiry. Each new piece of fibre is added to the rope for a good, justified reason. The underdetermination of theories by data is therefore less serious than sometimes claimed; theories are considered in their disciplinary context, their purposes of enquiry.[440] I will illustrate these points with a more detailed story of continuity through change in a scientific theory.

Michael Friedman describes a case where he wants to show the relation between later theories and Kant's physics, but he also in effect shows something that is more interesting for my current purposes. The case gives us an example of continuity *ex post*, complemented with the reasons for the choices, the factors that informed the decision-makers.

When detecting the historical roots of Einstein's theory in Kant, Friedman is not saying that the results were predictable. "There can be no question, of course, of Kant having 'anticipated' this theory in any way."[441] Instead,

> Kant's own conception of the relationship between geometry and physics…set in motion a remarkable series of successive reconceptualizations of this relationship (in light of profound discoveries in both pure mathematics and the empirical basis of mathematical physics) that finally eventuated in Einstein's theory.[442]

Friedman describes a chain of physicists and mathematicians in between Kant and Einstein, and how their positioning within the tradition makes sense given what they want to achieve. Continuing in this tradition, picking up elements from others, Einstein's theory represents "a natural (but also entirely unexpected) extension or continuation".[443] Friedman's analysis of an intellectual trail from Kant to Einstein for concepts describing the relation between geometry and physics illustrates how involved, and how *case-specific*, a historical account for the development of a theory is likely to be. One of Friedman's examples is Hermann von Helmholtz, who chose a generalisation of the Kantian spatial intuition that was "the *minimal* (and in this sense unique) such generalization consistent with the nineteenth-century discovery of non-Euclidian

[440] See my §4.7.
[441] Friedman [2009], p.266.
[442] Friedman [2009], p.266.
[443] Friedman [2009], p.265.

geometries."[444] Another is Ernst Mach, "who first forged a connection between Kant's original solution to the problem of 'absolute space' and the late nineteenth-century solution".[445]

Albert Einstein, a third example, is described as "delicately situating himself *between* Helmholtz and Henri Poincaré."[446] Einstein finds a "radically new way of reconfiguring the relationship between the foundations of geometry and the relativity of motion",[447] Friedman writes, but Einstein does so by accepting or rejecting different pieces of the respective approaches taken by Poincaré and von Helmholtz. Einstein makes his choices based on what is useful for his purpose, that is, on what would fit with the development of his ideas about relativity and gravitation. Pieces that do not fit represent theoretical dead-ends, without continuity. For example: "Poincaré's rigid hierarchy of the sciences…stands in the way of the radical new innovations Einstein himself proposes to introduce."[448]

Friedman discusses in detail how these scientists connect with their predecessors, to meet their own purposes. This is an extract of how he analyses Einstein's development.

> But why was it necessary, after all, for Einstein to engage in this delicate dance between Helmholtz and Poincaré? The crucial point is that Einstein thereby arrived at a radically new conception of the relationship between the foundations of (physical) geometry and the relativity of space and motion. These two problems, as we have seen, were closely connected in Kant, but they then split apart and were pursued independently in Helmholtz and Mach…In Poincaré…the two were perceptively reconnected once again…Indeed, it is for precisely this reason, as we now see, that Poincaré's scientific epistemology was so important to Einstein. Einstein could not simply rest content with Helmholtz's 'empiricist' conception of geometry, because the most important problem with which he was now faced was to connect the foundations of geometry with the relativity of motion.

[444] Friedman [2009], p.257.
[445] Friedman [2009], p.259.
[446] Friedman [2009], p.264.
[447] Friedman [2009], p.265.
[448] Friedman [2009], p.265.

> But Einstein could not rest content with Poincaré's conception either, because his new models of gravitation had suggested that geometry has genuine physical content.[449]

Reading this story, we do not get the impression that decisions could have gone "either way". There is change and discontinuity, but there is continuity too, going from Kant to Einstein; each step is taken for a good reason. This reason, however, was not due to a general criterion applicable throughout all science, but a specific one for a specific situation.

Ex post explanations have sometimes been accused of rationalisation and triviality. Chakravartty regards it as a challenge "to identify precisely those aspects of theories that are required for their success, in a way that is objective or principled enough to withstand the charge that realists are merely rationalizing *post hoc*".[450] The phrase "merely rationalizing *post hoc*" suggests that realists might take examples from the history of science and superimpose patterns that had in fact no role in the scientific development, just to make their hypothesis come out right. Psillos admits that "the whole idea of the specification of a core description involves an element of *rational reconstruction*…the reader may worry as to whether this reconstruction is ad hoc."[451] Nevertheless, Psillos defends his use of core descriptions, and claims that such a reconstruction does not need to be *ad hoc*. I believe he is right.

In the examples above, there is no question of an illicit, *ad hoc* rationalisation where we rewrite the past to suit our new theory; my point is not that we always can *describe* history in terms of continuous progress. Friedman's approach to the analysis of 'space' is based on historical examples, capturing the decisions that actually led science forward, and the process during which these decisions were taken. When we look back, using the *ex post* explanation, we can see the path scientific development took.

This gives me the response I have been looking for to the first part of Kuhn's Challenge, my point (A): *If the meaning of key terms change between theories on either side of a paradigm*

[449] Friedman [2009], p 265.
[450] Chakravartty [2017], §2.3.
[451] Psillos [1999], p.297. His "core properties" corresponds to what I have called "essences".

shift, how can we say that these theories are about the same thing? We *can* say that, is the answer, because there is an *ex post* story to tell about the continuity between the earlier and the later version of the term, a story describing the decision-making in the relevant scientific community, justifying continuity.

6.6 Conclusions

I this chapter, I have endorsed LaPorte's view that decision-making has played an important role in the history of natural kind terms. But I have also stressed that such decisions are not arbitrary. They are taken for good reasons in their respective disciplinary context, even if those reasons are only available *ex post*, from the perspective of the new theory. I also defended this model against suspicions of rationalisation. *Ex post* does not imply *ad hoc*. Recognising the crucial role of decision-making in the history of sciences, which Quine's Dictum tells us to do, does not threaten conceptual continuity and scientific realism when the role of good reasons also is recognised.

According to Hacking, what I have now described is not so far from Kuhn's position. In "Objectivity, Value Judgement, and Theory Choice"[452] Kuhn returns to the issue of theory selections, and identifies five general values.[453] But importantly, as Hacking stresses:

> [H]is five values [Kuhn believes] and others of the same sort are never sufficient to make a decisive choice among competing theories. Other qualities of judgement come into play, qualities for which there could, in principle, be no formal algorithm.[454]

If Hacking's reading is correct, Kuhn does not hold that decisions were arbitrary, nor that they could have gone either way, but that some crucial components for those decisions were qualitative, beyond formalisation and generalisation. I have made the point in terms of the

[452] In Kuhn [1977], chapter 13, pp. 320-339. The text was delivered to a conference in 1973.
[453] Summarised in Hacking [1983], p.13: "Theories should be accurate, that is, by and large fit existing experimental data. They should be both internally consistent and consistent with other accepted theories. They should be broad in scope and rich in consequences. They should be simple in structure, organizing facts in an intelligible way. They should be fruitful, disclosing new events, new techniques, new relationships."
[454] Hacking [1983], p.13.

weighting of relevant factors rather an in terms of qualitative components, but the conclusion is the same: Decisions are taken for a reason, but no general formula can predict or explain decisions taken when theories change.

This put me in a situation to revisit Putnam's historical-chain based argument for conceptual continuity. The argument from referential continuity can be strengthened by incorporating the crucial role of decision-making for good reasons.

In the beginning of the last chapter, I listed four criteria that a justification for scientific realism would have to meet:

(i) The justification must cover natural kinds and individuals.

(ii) The justification must be available to us, at least in principle.

(iii) It must make scientific theories fallible.

(iv) It must sort historic posits in the right way, separating referring from non-referring historical terms.

The solution proposed meets these criteria and is aligned with my methodological rule, Quine's Dictum, which says that if something is an established part of science, "we cannot ask better than that". This is the type of continuity we *can have*, a case-specific *ex post* type of continuity, mirroring actual intellectual progress, where choices are made based on what makes sense for a given intellectual enterprise, given empirical progress. Is there anything missing? This account does not provide general *ex ante* guidelines for continuity and progress. But insistence on general, *ex ante* rules are left-overs that should be given up together with the PTT.

7. The Semantics of Natural Kind Terms

7.1. Introduction

Some philosophers that I have quoted and agreed with in earlier chapters, including Ghiselin, LaPorte and Chang, have expressed scepticism regarding the compatibility of their conclusions with the Kripke-Putnam semantics that I outlined in Chapters 2 and 3. In opposition to that scepticism, this chapter claims that many (but not all) insights in the Kripke-Putnam semantics can be combined with points raised by some of their opponents, and with the framework developed in this thesis, to point toward a more satisfactory semantics for natural kind terms.

Using the framework, I defend Kripke's claim that natural kind terms "have a greater kinship with proper names than is generally realized."[455] I also defend Putnam's second argument for conceptual continuity, the necessity-based continuity argument, and claim that this argument is sound, if suitably supplemented. In addition, I return to what is missing in Godfrey-Smith's definition of scientific realism. To do that, I first need to introduce another piece of machinery, the global commitments of science, which tells us why natural kind terms are rigid.

7.2. Rigidity and Scientific Identifications

In Chapter 2, I presented Kripke's theory of proper names against the background of the descriptionist theory. I also presented the more briefly described and more controversial extension of his semantics to natural kind terms. Kripke argues that natural kind terms are very similar to proper names and that scientific identifications are very similar to identity statements with proper names. But this is not obvious, and on the face of it there are important differences. I earlier mentioned two worries: the first that natural kind terms are not rigid at all, and secondly that if they are, this rigidity has another explanation than in the proper names case, making the parallel less close and the extension less credible. I have already discussed the first point,[456] but I will now need to address the second, which was left hanging until I could develop some machinery to resolve the issue. I will claim that both Kripke and his critics have good points.

[455] Kripke [1981], p.134.
[456] See §2.7.

The rigidity of proper names follows our natural understanding of how we use names, as tags that follow an object through time and possible worlds.[457] That necessity of identity statements follows from rigidity is also intuitively clear: if two names always refer to the same object, variation in time and possible worlds make no difference to the truth of the identity statement. But the scientific identifications appear to be different, because rigidity of the terms does not seem to be what guarantees the necessity of the statement. Scientific identifications do not *prima facie* look like employing two names for the same object; they look like they assign properties, necessary or essential properties.[458] The term 'H_2O', at least on a natural reading, is not a name; it is an abbreviated description of a chemical compound. But 'water is H_2O' differs also from 'water is what fills our oceans', which also assigns a property to water. Water can exist even if the oceans dry out, but not without being H_2O; being H_2O is an *essential* property of water while filling the oceans is a contingent one. Scientific identifications assign essential properties to natural kinds, and must be necessarily true. There is an asymmetry: being H_2O is the essence of water, not the other way around, even if being water probably is a necessary property of H_2O. This asymmetry is not present in 'Hesperus is Phosphorus', so the cases appear different.

Penelope Mackie has raised another objection to the Kripke-Putnam account of scientific identifications. This objection is in a sense the opposite of the one I have just outlined. The theoretical identification 'water is H_2O' should according to the Kripke-Putnam semantics be understood to mean 'It is necessarily true that a sample consists of H_2O if and only if it consists of water'. But for Mackie, this example "does not, on the face of it, attribute an essential property to anything. Yet Putnam's view is usually described as a version of essentialism."[459] Mackie maintains that for an essentialist, theoretical identifications are assignments of essential properties. She has in mind a typical expression being: "❑ (a exists -> a is F)",[460] where the attribution of the essential property F is clear.

[457]'Tag' was the term used for this purpose by Ruth Barcan Marcus [1961]. It roughly corresponds to rigidity. It has the advantage of being a more intuitive term, but the disadvantage that it excludes connotation altogether.
[458] Kripke uses 'necessary properties' and 'essential properties' as synonyms. For an argument that they should not be, see Fine [1994].
[459] Mackie [2006], p.12,
[460] Mackie [2006], p.6.

Mackie moves on to discuss modifications of her definition of essentialism to accommodate how she reads Kripke and Putnam. My claim in this section is instead that we can accept essentialism for natural kinds in Mackie's (unmodified) sense while still holding that natural kind terms are rigid, in the *same* sense as proper names. This way, a still stronger case for the extension of the Kripke-Putnam semantics from proper names to natural kind terms than I have built so far, with an even closer parallelism, can be constructed.

To show this, I need to recall Kripke's distinction between the baptiser who introduces a term and subsequent users.[461] The baptiser fixes the reference of a natural kind term to an abstract object, the natural kind. Subsequent users use the term in the same way as previous users or (in Putnam's improved version) as it is used in their linguistic community. The first issue above, the asymmetry of scientific identifications that I raised, relates to the baptism situation. The second, the *lack* of essentialist attribution that Mackie points to, applies to subsequent usage.

It is indeed plausible that the statement 'water is H_2O' at one point in time, namely during the initial baptism, represented an assignment of an essential property, implementing a scientific discovery, which became a part of the reigning paradigm. This follows a process where a theory is accepted which states that the essence of water is its microstructure. In the dubbing (or redubbing) taking place, an abstract object is baptised, perhaps using an old term with a modified meaning. At this stage, the theoretical identification 'water is H_2O' is (as Aristotle says) best read as a definition.[462]

In this view, the initial baptism is an assignment of an essential, defining property to an abstract object. But this is surely not how we use a natural kind term later on. Having accepted the scientific identification, later references to 'water' and 'H_2O' function not unlike two different names for the same (abstract) object, no longer as a definition, because water can no more appear without H_2O than Phosphorus can appear without Hesperus, and vice versa.

[461] See §2.5.
[462] Aristotle *Metaphysics,* 1031 a12: "[D]efinition is the formula of the essence."

As name-equivalents, both terms are rigid; they name the same object in all possible worlds. The scientific identification 'water is H_2O' is necessarily true, in virtue of this rigidity. Therefore, in normal use, the rigidity notion functions in the same way for natural kind terms as for proper names.[463] Theoretical identifications, after the initial reference setting, represent a true identity statement.

If we regard natural kinds as abstract objects to which the transworld identity assumption applies, that have their essences set in an initial baptism, the similarity between proper names and natural kind terms is very close – as Kripke said.

7.3. Global Commitments

I have followed Hasok Chang in his defence of pluralism, but I disagree with one of his conclusions. Chang accepts that that 'water is H_2O' expresses an objective truth. "However, this truth is internal to various systems of practice in which it is true."[464] There are, Chang shows, other systems of practice, yielding different truths. He continues: "[T]ruth as I conceive it means correctness as judged within a specific system of practice".[465] Chang's conclusion might seem to follow naturally from pluralism, but I believe that this kind of antirealism can be avoided, in favour of a view closer to our realist intuitions, if we take onboard an insight (briefly) raised by Kuhn in *Structure* and a valuable part of Putnam's realism in "MoM". I will in this section describe what I mean by "global commitments" and how this explains the rigidity of natural kind terms. In the next section I claim that global commitments can be used to underpin Putnam's necessity-based argument for conceptual continuity.

In "MoM", Putnam points to the *semantic* connection between realism and scientific practices, and this is the part that should be saved.

[463] Salmon [2003] briefly mentions a similar idea.
[464] Chang [2014], p.214.
[465] Chang [2014], p.214.

> [F]or a strong antirealist *truth* makes no sense except as an intra-theoretic notion...he does not have the notions of truth and reference available *extra-theoretically*. But *extension is tied to the notion of truth*. The extension of a term is just what the term is *true of*.[466]

And he adds:

> My point is that if we are to use the notions of truth and extensions in an extra-theoretic way (i.e. to regard those notions as defined for statements couched in the languages of theories other than our own), then we should accept the realist perspective to which those notions belong.[467]

We note that Putnam in these passages does not directly argue for realism (as he does elsewhere[468]). The subtitle "Let's be realists" used in "MoM" might show an attitude or a *commitment* to realism rather than an argument in favour of it, an attitude that comes with doing science. What Putnam says in this quote is that if we want to use the notions of truth and extension extra-theoretically, across paradigm shifts, realism comes in the package. But this does not mean that perfect-theory realism follows automatically, I concluded in Chapter 5. Putnam's semantic point about the use of 'truth' and 'extension' is, contrary to what he believes, equally compatible with a pluralistic view of realism. Putnam's mistake, both as a realist and as an antirealist, is that he gets his options wrong.[469]

To Putnam's point about scientific commitments being "global", I will now add a point from Kuhn, where he tells us about the *nature* of the scientists' commitments.[470] During normal science, Kuhn says, the scientists in their respective communities are committed to a paradigm; otherwise they would cease to be scientists. He regards paradigms as implying a "strong network of commitments – conceptual, theoretical, instrumental and methodological".[471] That network of commitments "must extend to areas and to degrees of precision for which there is no full

[466] Putnam [1975], p.236.
[467] Putnam [1975], p.237.
[468] See my Chapter 5.
[469] I suspect that Kuhn is guilty of the same mistake, see my §1.8.
[470] With the intended sense, 'global' is a close relative to Quine's 'cosmic' and Putnam's 'extra-theoretical'.
[471] Kuhn [2012], p.42.

precedent".[472] This is what generates the puzzles on the agenda of scientists doing normal science. But Kuhn also recognizes *other* types of commitments, at *higher levels* than paradigms. One level contains commitments that are similar to paradigms, but on a more abstract level, such as a commitment to Descartes's corpuscles theory, which influenced many actual paradigms, as have commitments to principlism or compositionism.[473] A third type is on a still higher level, and even more abstract. This third level consists of mandatory requirement for all scientists, for example the commitment "to be concerned to understand the world and to extend the precision and scope with which it has been ordered".[474]

For Kuhn, paradigms consist of theories and methods. He is well known for his description of how paradigms form worldviews and intellectual universes for scientists. But in the passages I quoted above, Kuhn also recognises guiding principles *across* paradigms, functioning as meta-paradigms, including fundamental claims and commitments of science. The nature of these commitments by scientific communities is one of several areas that are underdeveloped in *Structure*. One important feature missing, I suggest, is the global claims that Putnam picks up in "MoM". If we add this component, we can say that a theory, in virtue of being a *scientific* theory, is understood to apply to all previous cases and to all future cases, and also to types not yet encountered.[475] One example in *Structure* is Copernicus' prediction of Earth-like properties for other bodies in the solar system.[476] The claim is global, applicable across time and possible worlds.[477]

7.4. Necessity-Based Continuity

The global commitments of science, outlined in the previous section, supplement Putnam's second argument for conceptual continuity, the necessity-based argument. They also give us an

[472] Kuhn [2012], pp.100-101.
[473] See §6.3.3.
[474] Kuhn [2012], p.42.
[475] Note that the notion of global commitments just mentioned is not identical to the idea that individual scientists have been scientific realists, and that these beliefs are relevant in explaining the success of their endeavours. That idea is rightly criticised by Laudan [1981] and by Devitt [1983].
[476] See §1.4.
[477] Depending which modal categories of necessity one recognises, "possible worlds" here might need the qualifier "nomological". I will ignore this complication.

answer to the question *why* natural kind terms are rigid, and the information needed to see what is missing in Godfrey-Smith's definition of scientific realism.

I will first recall the issue. I earlier wrote that the rigidity of 'water' and 'H_2O' can be used for one type of continuity argument, the necessity-based argument.

> The scientific identification 'water is H_2O' is necessarily true because it is true in all possible situations, due to the rigidity of the terms involved. Consequently, it is true over time.[478]

But this is not by itself an argument for conceptual continuity either, I found, as it takes the continuity of objects over time and possible worlds for granted. The issue is the scope of the rigidity of natural kind terms, and the necessity of scientific identifications. The argument uses an extra premise, an essentialist assumption, to support extra-theoretical conceptual continuity, without which we cannot justify realism.

What matters for conceptual continuity, according to the idea I am presenting, is not the beliefs of individuals (dead or alive), but the commitments built into the community of practices, and the meta-paradigms that guide them. The signs of those commitments can be seen in the semantics – the rigidity – of the scientific natural kind terms, as Kripke and Putnam show us. The global commitment of scientific practices gives the explanation and the basis for the semantics; the rigidity reflects the global ambition of typical scientific entities and regularities. When we make the initial baptising of a natural kind object, with a scientific identification, we are at the same time saying that this identification is globally true, in all circumstances, across time and possible worlds. But while scientific practices contain *this* type of commitment, there is no commitment to the PTT; the multitude of incompatible theories and models indicates quite the opposite.

It is true that the global commitment to scientific identifications such as 'water is H_2O' is made using 'water' to mean '$water_{Today}$'. But this is what we must do when following Quine's Dictum:

[478] §3.2.1.

we must refer to the methods used by the current sciences, which can motivate the higher explanatory power of current sciences compared to their predecessors. Our sciences do not guarantee truth, but they do give us fallible justifications for our beliefs in them.

In the last chapter, I quoted Godfrey-Smith's definition of scientific realism:

> One actual and reasonable aim of science is to give us accurate descriptions (and other representations) of what reality is like. The project includes giving us accurate representations of aspects of reality that are unobservable.[479]

But when I discussed this, I also said that something important is missing from Godfrey-Smith's definition. We can now pinpoint what is missing: it is the sciences' commitment to *global* answers, which results in natural kind terms being rigid and scientific identifications necessarily true, "given our language".

7.5. Rigidity, Necessity and Temporal Indices.

Having agreed with Kripke's view about the rigidity of natural kind terms and the similarity between identity statements with proper names and scientific identifications, and after having introduced the role of global commitments in science, I will in this section discuss the following questions:

- If scientific identifications are true only when the natural kind terms have temporal indices, and if changes of indices might include meaning changes, does this mean that they are not necessarily true?
- If natural kind terms are rigid only with temporal indices, does that rule out conceptual continuity?
- Is there a place for the necessary *a posteriori*, which is so important for Kripke?

I will start with the first of these questions, about necessity and meaning changes.

[479] Godfrey-Smith [2003], p.176.

For Kripke and Putnam scientific identifications are necessarily true if true at all, that is they are true over time and across possible worlds. For Kripke, this necessity exists given the use of terms with today's meaning, while Putnam adds that the meaning has not changed substantially over time. He argues for this in his thought experiment about Archimedes and 'gold'. Archimedes, Putnam claims, must have had approximately the same opinion of the essence of gold as we have, in effect the same opinion on its essence type, otherwise 'gold is chemical number 79' could not be true in ancient Greece, and therefore not necessarily true at all.[480] But this is not the case, so long as we keep our indices in order and realise that 'gold is chemical number 79' is necessarily true (so it is true also for Archimedes' time) if 'gold' means '$\text{gold}_{\text{Today}}$' as it does for us. Earlier meanings of 'gold' come into the picture when we discuss continuity over paradigm shifts, but not when we discuss necessity. We have been convinced by Kripke's and Putnam's thought experiment that gold always has been chemical number 79, and mix up this question with the issue of what Archimedes and his contemporaries meant by 'gold'.

Our intuitions might be different when we turn to projection of scientific identifications into the future. If 'gold is chemical number 79' is necessarily true, it must have been true in Archimedes' days, and it must also be true forever. Because this is a very strong condition, one might be tempted to give up the global ambitions of necessity after all. But that would be another example of a misunderstanding of the commitments implied by scientific theories. It is in the logic of scientific regularities that they are assumed to be truly global, with rigid terms, and apply over time, both backwards and forwards, and across possible worlds. The identity statements might later on be changed, but this does not take away the global commitments that comes with science, reflected in the necessity of scientific identifications. These are not dependent on science stopping here and now.

The Kripke-Putnam semantics allows a fully competent speaker to make mistakes and to be ignorant regarding an object's necessary properties. The object can still be correctly identified, based on properties it has contingently. But what is contingent and what is necessary? My framework says that this is theory-dependent. We can (plausibly) imagine a situation where the

[480] See §3.2.

essential nature of $water_{1150}$ was functional. If so, people living before scientific discoveries were ignorant about the microstructural nature of water, but it is an ignorance that matters from *our* perspective, where water has a microstructural essence, not from a 1150 perspective, when it did not.

Let us look at the temporal indices again.

'$Water_{1150}$ is H_2O' cannot be necessarily true, or indeed true at all, as this statement makes no sense at any time. This is a unicorn situation: we cannot formulate the conditions under which that sentence would be true.[481]

The situation in 1750, when the Twin Earth thought experiment takes place, is more complex. Speaking in 1750, '$water_{1750}$ is H_2O' is not yet a necessary truth, although it was known (or could be known) that the identity of water with a specific chemical compound would be necessary *if discovered*. An assignment of an essence type has taken place, with a variable placeholder for an essence, but it is tentative. But these meaning changes do not affect the fact that '$water_{Today}$ is H_2O' is necessarily true. I will expand on the issue of continuity vs. change in section 7.6.

I will now turn to the second question, the question of whether rigidity with temporal indices is incompatible with conceptual continuity. I answer "no", but with the caveat that extra-theoretical continuity of natural kind terms is not guaranteed by their rigidity.

I earlier quoted Ghiselin pointing to the dynamic nature of natural kind terms in science.[482] I will now also quote the continuation:

> [S]cientists do not attach a name to a class, then discover the defining properties which are its essence, but rather redefine our terms as knowledge advances. Therefore the view of Kripke…that natural kind terms are, like proper names, 'rigid designators', should be dismissed as nugatory, and with it the accompanying essentialism.

[481] See §3.4. and Kripke [1981], p.24.
[482] In §4.2.

But Ghiselin's conclusion about the view that scientific natural kind terms are rigid ("nugatory") is not convincing; it does not follow from the fact that the meaning of terms can change that they are not pointing to the same kind in all possible worlds *at a given time with a given meaning*.[483] His criticism of essentialism is valid against the strong version of essentialism, such as the one expressed in "MoM", which includes an extra-theoretical claim, but it is not valid against an essentialism without such claims. It is the latter version that I use in this thesis.[484]

LaPorte also denies that 'water is H_2O' is necessarily true.

> As far as we know, it is metaphysically possible that we could find a substance revealing more open texture, so we do not know that 'water' and 'H_2O' refer straightforwardly to the same items in all possible worlds.[485]

However, this problem is resolved with indices: the fact that H_2O might not be the essence of "$water_{2075}$" does not affect the fact that it is the essence of "$water_{Today}$".

The semantic property of some terms to be rigid generates necessary truths. Rigidity says something about counterfactual possible worlds. As a special case, it also operates across time. The rigidity of natural kind terms is underpinned by the essences of the kinds, assumed by the global, universal commitment of science: to explain what there is, what has been and what will be. As I argued in the previous chapter, following Kuhn, this commitment is backed up by the relative superiority in explanatory power of subsequent theories, *not* by any theory being closer to a perfect theory.

Rigidity over time and possible worlds must not be confused with extra-theoretical continuity: the former is a feature of the meaning of natural kind terms with their current meaning; the latter

[483] LaPorte [2013], p.57, writes about words that have changed meaning that "[w]e should not expect such a word to designate rigidly the *same* condition now as it did in earlier times. But *at any given time* it could still rigidly designate a single condition".
[484] I define an essence as the sufficient and necessary conditions determining a natural kind and giving it explanatory power. The role as a carrier of extra-theoretical continuity is an extra, alleged property of essences, and not a part of my definition.
[485] LaPorte [2004], p.110.

is a relation between terms with their current meaning and related terms with a previous or later meaning, which can only be specified *ex post*.

This analysis has implications for the conclusions we can draw from rigidity. Putnam writes that an operational definition will not do for natural kind terms. "Rather 'we use the name *rigidly*' to refer to whatever things share the *nature* that things satisfying the description normally possess."[486] But my point here is that the fact that we use the terms rigidly, applying the "Today" index, does not justify drawing conclusions about terms with indices for earlier periods, conclusions about what people living long before us might have meant. In Chapter 2, I quoted an old riddle about a horse's legs and tails, used by Hughes as a characterisation of Kripke's view.[487] There is a similarity between this riddle and Putnam's Twin Earth argument, as both trade on an ambiguity regarding which language we are using. Likewise, Putnam establishes, with his thought experiment, that given that we by 'water' mean '$water_{Today}$' then 'water is H_2O' is necessarily true. *A fortiori*, it follows that 'water is H_2O' is true for the situation in 1750 and in 1150. But the meaning of 'water' in this statement is '$water_{Today}$' and Putnam has not showed us that '$water_{Today}$' has the same meaning as '$water_{1150}$'. The $same_{Liquid}$ relation does not help, as this immediately introduces the assumption that water has a chemical as its microstructural essence, the legitimacy of which, before the birth of modern chemistry (or after it has been replaced) is in question. The Twin Earth thought experiment establishes the rigidity of the terms – but not the extra-theoretical continuity over time.

Giving up the ambition to make rigidity explain meaning constancy is not saying that Archimedes could not talk about gold, tigers or water. It *is* saying that this is not guaranteed by a semantic theory about reference. I have presented two ways conceptual continuity can be defended, both based on Putnam's "MoM" arguments, suitably complemented. The first is the necessity of 'water is H_2O', backed up by the global commitments and improving methods of

[486] Putnam [1975], p.238.
[487] "If 'leg' meant 'tail-or-leg', how many legs would a horse have?" "Five." "No, four: calling a tail a leg doesn't make it so."

science. The second is the historical links of reference, backed up by decision-making during paradigm shifts, where the meaning constancy of some terms, for good reasons, has been upheld.

Finally I turn to the necessary *a posteriori,* which plays an important role in *N&N*, making it possible for Kripke to explain how scientific identifications can be necessary although unknown. His argument looks difficult to combine with an analysis similar to mine. Indeed, LaPorte states that while philosophers "generally acknowledge necessity, and in particular a posteriori necessity...[t]he combination is confused."[488]

This is the issue. If we first look at '$water_{Today}$ is H_2O', I have argued that this is necessarily true, backed up by global scientific commitments. But it does not seem to be *a posteriori*. The vernacular term is influenced by the scientific definition, and refers to the same natural kind. It is true that Kripke regard knowledge *a priori/a posteriori* as applying to individuals, and it is of course also true that some people even today do not have *a priori* knowledge of the identity. But this does not help to save *a posteriori* status if we also accept Putnam's convincing notion of the linguistic division of labour. The fact that the knowledge most definitely exists among chemists today, requiring no further empirical studies, means that it exists in the linguistic community too, backing up current everyday usage.[489] An individual can still can be mistaken, or fail to consult the best knowledge available in his reasoning. But rather than an *a posteriori* situation, this resembles Kant's non-pure *a priori* in the story about the careless builder.

> Thus we would say of a man who undermined the foundation of his house, that he might have known a priori that it would fall, that is, that he need not have waited for the experience of its actually falling. But still he could not know this completely a priori. For he had first to learn from experience that bodies are heavy, and therefore fall when their supports are withdrawn.[490]

[488] LaPorte [2004], p.165. See also Jubien [2009], chapter 7, and Chenyang Li [1993].
[489] See §4.3.
[490] Kant [1933], "Introduction", p.43, B3.

If we turn to 'water$_{1750}$ is H_2O', it is *a posteriori* – but it does not seem to be necessary, as in 1750, the discovery has not yet happened, and the decision not been taken. There is no commitment yet from the scientists.[491] We could say that the identity between water and H_2O was necessary and *a posteriori* for people in 1750, but from our perspective today, using our language. It would in a sense be correct. But this means falling back on the riddler's old tricks again: we would be saying that the scientific identification is necessary when 'water' means 'water$_{Today}$' and *a posteriori* when it means 'water$_{1750}$'.

In short, the necessary *a posteriori* in the *N&N* sense does not fit easily within my framework. Fortunately, it is not needed either. There is no problem to explain the fact that a statement that is necessarily true today was *not* necessarily true before a redubbing took place, if rigidity and necessity are based on commitments to the global application of our best theories and natural kind terms appropriately indexed.

One way to look at this conclusion is to see it as the result of the remaining major difference between the reference of proper names and the reference of natural kind terms. The former refer to physical individuals where the transworld identity assumption is very strong.[492] The identity is necessary over time and across possible worlds, undisturbed by whether we know about it or not ("we are talking about *him"*). The intuition behind the necessary *a posteriori* is clear. The latter, in contrast, refer to abstract objects, defined by an essence, within an essence type, subject to review at paradigm shifts, where the necessity of scientific identifications does not always remain unchanged in a redubbing. It is not obvious that the necessary *a posteriori* has any role to play in this context.

7.6. Continuity vs. Change

Here is what looks like a dilemma. On the one hand, it sounds right and even obvious to say that water has always been H_2O; it was H_2O in the middle ages, in antiquity, and long before that. A

[491] See §§7.3. and 7.4.
[492] See §3.4.

series of thought experiments by Kripke, Putnam and others certainly mobilise our linguistic intuitions in that direction. Godfrey-Smith formulates this first intuition in the following way:

> Unless we have made some very surprising mistakes in our current science, the world we now live in is a world of electrons, chemical elements, and genes, among other things. Was the world of one thousand years ago a world of electrons, chemical elements, and genes? Yes, although nobody knew it back then.[493]

On the other hand, it is *also* sounds right to say that kinds like water come with modern chemistry, when their essences are defined. No essence can be defined without an essence type, and to choose a microstructural essence type was not trivial, I have argued. It took a theory and a decision. Godfrey-Smith illustrates also the other leg of the dilemma, the opposite intuition, when he writes about electrons that

> the concept of an electron is the *product* of debates and experiments that took place in a specific historical context. If somebody said the word 'electron' in 1000 A.D., it would have meant nothing – or at least certainly not what it means now. So how can we say that the world of 1000 A.D. was a world *of* electrons? We cannot; we must instead regard the existence of electrons as dependent on our conceptualization of the world.[494]

I agree with Godfrey-Smith that there are good arguments for both positions, and I think a realist has to account for both. But note that the dilemma only arises if we think we have to *choose* between the two. According to my framework, we do not.

We can find the same issue in a comment by Magnus on an idea by Richard Boyd. Boyd holds that samples of water or oxygen have always existed but that the natural *kinds* water and oxygen came along with the chemical revolution. Magnus objects, in line with Godfrey-Smith's first intuition above: "Of course the ancients had no word for it, I would say, but **oxygen** was around then."[495]

[493] Godfrey-Smith [2003], p.173.
[494] Godfrey-Smith [2003], p.173.
[495] Magnus [2012], p.107.

I believe both that it is true that water has always been H_2O *and* that 'water' has changed its meaning as a result of changes in scientific theories. To show that, I will first have to recall some distinctions made in earlier chapters, and then apply the temporal indices to the Twin Earth case.

We have the following milestones. At one point, naturally connected to the Chemical Revolution, a paradigm is accepted, on the basis of which a hypothesis is formed that water is a scientific natural kind in chemistry, that is that the essence type of water is a chemical, a microstructural essence type. In 1750 (I assume), it had been accepted that solids and liquids have an underlying microstructure, and that unique superficial properties normally correspond to a unique microstructure. The paradigm has been established. In the phase of normal science that followed, the details are provided, and the exact chemical compound identified.[496] But the identity between a natural kind identified by its observational properties and a unique microstructure could only be regarded as a hypothesis, and such hypotheses do not always work out. In 1750, we cannot know if one, two or many microstructures will be found, and we cannot know how the discovery will be handled.

In §4.4 I mentioned that Kripke offers a potential model for the transfer from the situation for 'water' in 1750 to the mature science phase. Discussing species, he says: "[S]cientific discoveries of…essence do not constitute a 'change of meaning'; the possibility of such discoveries was part of the original enterprise."[497] Maybe I can paraphrase his idea in this way: the meaning is fixed by the decision on essence type, like an open statement with a variable, or a slot later to be filled in. Changing Kripke's example, a reasonable interpretation is that the original enterprise in the case of 'water' took off in 1750.

Because we know the outcome for 'water' – that is, because we know the meaning of '$\text{water}_{\text{Today}}$' – we can describe the situation in 1750 as one where a gap needed to be filled by identifying the postulated microstructure. But this is not an accurate description of the real 1750 situation, as it does not represent the options that in principle existed at the time. Put another way, the story that

[496] Within this paradigm, on the other hand, there was no place for questions about how water constituted the basis for other substances to form the world , which occupied Thales' thoughts.
[497] Kripke [1981], p.138.

is (turned out to be) true of water, its identity with a unique chemical microstructure, cannot be generalised to all cases, for reasons given; the result can differ from the hypothesis, and in a crisis, several options are open.

Maybe this point is better illustrated if we again take the time machine in the opposite direction. It is conceivable that science in 2075 will discover that the formula H_2O in fact covers two radically different configurations on an underlying, even more basic level, previously established and now part of chemistry$_{2075}$. What will we then do with 'water'? Like 'jade' we could decide to continue to call both substances 'water', or like 'ruby' decide to separate the two, or like 'reptiles' give up using the term for scientific purposes.[498] We do not even know which factors will be the relevant ones, and what will be their weighting: will this be based on chemical classification only, or will historical, functional or commercial considerations play a role? Also, the 2075 discovery just described might not be possible within the current theories of chemistry and physics. It would therefore trigger a major scientific crisis, which eventual outcome, the new paradigm, we cannot possibly know today.

Empirical discoveries are highly relevant. An assumption of a microstructural essence can fail, or in more general terms, the choice of an essence type might be altered in the light of such discoveries. This fact introduces a timeline issue about meaning.

For all they knew in 1750, the hypothesis that water is H_2O could have turned out to be wrong, and Putnam discusses this scenario. "But the local water…may have two or more hidden structures – or so many that 'hidden structure' becomes irrelevant, and superficial characteristics become the decisive ones."[499] Note Putnam's use of verb in the phrase "*become* the decisive ones" [italics added]. As the essence determines the meaning, he has to say that the discovery of all the microstructures would affect the meaning of 'water'.

Choices of natural kinds depend on the purpose of enquiry. Had Putnam's scenario with an impractical number of microstructures occurred, one could imagine that 'water' would be retired

[498] We cannot have discover that water does not exist – but we can retire 'water' as a scientific term.
[499] Putnam [1975], p.241.

as a natural kind term in chemistry, as 'rodent' was in biology. It would of course continue to be a vernacular natural kind term for everyday purposes, and also on this point I agree with Putnam: the functional essence type based on observational properties would most likely be the relevant one.

Let us again call the actual world "W1" and the counterfactual world, with the multiple microstructures for water, "W2". Let us also stipulate that community of chemists in 1750 held the hypothesis that water has a chemical substance, without knowing which one, and that this issue was settled in 1780 for both worlds (and stable since). In W1 this was of course done with the acceptance that water is a compound that includes hydrogen and oxygen, and in W2 with the discovery that water lacks a common microstructure, and the subsequent acceptance of a functional essence based on observational properties.

With these scenarios, the meaning of $\text{water}_{\text{Today-W1}}$ is clearly different from the meaning of $\text{water}_{\text{Today-W2}}$. It also seems clear that '$\text{water}_{\text{1150-W1}}$' = '$\text{water}_{\text{1150-W2}}$'. The question is now which meaning changes took place over time in the two worlds, and when they took place. In my opinion, the most natural explanation is to say, *pace* Kripke, that a meaning-change in W1 took place in 1780, not in 1750, and that the meaning of 'water' in W2 remained unchanged.

In 1750, a hypothesis exists about the appropriate essence type, but the essence has not yet been identified, so there is insufficient information for a redubbing of the natural kind object. And as the nature of the scientific identification has not been formulated, it cannot yet influence the vernacular term.

I can now return to the threatening paradox, applying my indices. We have as before:

a. '$\text{Water}_{\text{Today}}$ is H_2O' is necessarily true.

This handles the first intuition; since it is necessarily true, it was always the case. I therefore disagree with the view Godfrey-Smith formulates when he asks: "So how can we say that the world of 1000 A.D. was a world *of* electrons? We cannot". I think, with Magnus, that we can confidently say that the world of 1000 AD *was* a world of electrons, chemical elements and

genes, as long as we are clear about which language we are using, which of course is our current one. I therefore *also* agree with Boyd, because **a.** does not directly entail:

> **b**. '$Water_{1000AD}$ is H_2O' is necessarily true

For **b.** we would need an additional argument, and our own linguistic intuitions, anchored in $language_{Today}$, are obviously *not* suitable for the job.

My confidence in **a**. is not dependent on **b**. being true, but rather on **c**. and **d**.:

> **c**. The claims of current, mature sciences are global, in the sense that they are valid over time and possible worlds.
>
> **d**. The methods of current, mature sciences are such that they justify (fallible) beliefs in the verisimilitude of their postulates, given a purpose area of enquiry.

7.7. Mistakes and Ignorance

I discuss a different pair of conflicting intuitions in this section: on the one hand Kripke's argument that we can be wrong about what we know about an object and still successfully refer, and on the other hand that if we use a term to talk about something lacking all the properties normally associated with the object, we have changed the subject.

Kripke sometimes seems to argue that a speaker might be wrong about every single fact he believes about an object and still successfully refer. "[W]e might…find out tigers had *none* of the properties by which we originally identified them."[500] One exception to that is however mentioned: if the fact in question is an essential property.[501]

We earlier found that a classification is needed before an essence can be identified, and I have called this classification for the choice of an essence type. But this is not enough to define a

[500] Kripke [1981], p.121.
[501] See Kripke [1981], p.14.

natural kind term; a worry remains about *what* we are defining and whether we can be so radically mistaken as Kripke thinks. Stanford and Kitcher write:

> It turns out that we can't in fact be wrong about *most* of the stereotypical features associated with a natural kind term. Suppose we identify having physical structure XYZ with qualifying for the reference of term T, but then discover objects with XYZ that have absolutely none of the stereotypical properties of the things to which we applied T. This can only count as discovering instances with none of the stereotypical features if we *refix* the reference of T through the description 'having physical structure XYZ.'[502]

This apparent conflict can be resolved with the help of the distinction between the baptiser's situation and that of subsequent users that I discussed in §7.2 above. The result of this analysis is that Kripke is right in some situations and that Stanford and Kitcher are right in others.

The *baptiser* cannot very well have been totally ignorant, or totally mistaken, as the eventual move from a pre-scientific to a scientific term shows, when a redubbing takes place. The abstract object referred to by the natural kind term might have lacked any of the observational properties associated with the pre-scientific term, but hardly *all* of them, in which case the term would be referring to some *other* object, or to none at all.[503] This also goes for a pre-scientific baptism: the baptiser must get important properties right. Subsequent *users* of the term, on the other hand, are using the term correctly if they are using it in the same way as in their linguistic community, which in extreme cases could be relying on descriptions of properties that the object in fact does not have.

Again, it can be objected that Kripke showed that it is possible to refer to an entity with descriptions of contingent properties only, and that a chain of reference makes sure that later mentioning refers to the same entity. But as we have seen, this argument only works with an essentialist assumption, the assumption that there is an entity whose essential/core properties

[502] Stanford and Kitcher [2000], p.111.
[503] We can compare with the multiple adjustment of the term 'acid' when Arrhenius, Brønsted and Lewis eliminated *some* criteria (sour taste and corrosiveness) to be able to give a microstructural explanation to the *other* observational criteria (see §4.3. in this thesis).

remain unchanged over time. Such a stipulation needs a justification, and this can be provided – but only *ex post* from the perspective of a later theory. A baptiser who in a redubbing situation tells us that he intends to use a particular known term to refer to an entity with none of the properties people at that time associate with the entity, is unavoidably going to be accused of missing the point, or of changing the subject.

7.8. Vagueness and Decision-Making

The distinctions between different temporal perspectives and between the related uses of language are also helpful to address a debate between LaPorte[504] and Alexander Bird.[505] LaPorte describes scenarios containing small versions of Kuhnian crises, leading up to an eventual outcome – the new theory being established – as in the cases of 'ruby', 'topaz' and 'jade'. His conclusion is that a decision is needed to reach this outcome and that the term is *vague* before the decision is taken. 'Water' before 1750 was vague as to whether XYZ was part of its extension, LaPorte claims; had it been discovered, it would have taken a decision, involving a genuine choice, to rule out XYZ out from the extension of 'water', and the outcome of that decision was not fully determined by the scientific facts. When the decision has been taken, the term has become more precise; after the Twin Earth encounter in 1750, and the subsequent decision that Putnam postulates, 'water' would no longer be vague in its relation to XYZ.

Discussing LaPorte's version of the Twin Earth, Bird makes two points. Firstly, he claims that 'water' before the Twin Earth encounter was *not* vague, and that we instead, in this thought experiment, see a change of meaning. Indeed, Bird argues that a new theory often brings a meaning change rather than a precisification of the old term, as LaPorte claims. Secondly, he argues that LaPorte's examples are incompletely described, and that the outcomes actually *were* determined by facts known beforehand, leaving no significant space for decision-making.[506] Bird's claim is that the authority on chemical substances is the society of chemists, and that there are good reasons from the point of view of chemistry to include D_2O in the extension of 'water'.

[504] LaPorte [2010].
[505] Bird [2010].
[506] Unless it is decided to *fundamentally* change the meaning of the term, Bird says.

LaPorte's version of the Twin Earth example is therefore not more realistic than Putnam's; it is open to the same objections.[507]

I am not sure Bird's first point hits LaPorte's real position. As I read LaPorte, his conceptual vagueness (sometimes called "open texture") only applies in a Kuhnian crisis situation, that is, in a situation where there is new data in response to which the old theory is inadequate.

Regarding the second point, Bird is right that the option to exclude D_2O from water that LaPorte's mentions in his version of the Twin Earth thought experiment, would not have made sense for chemistry. But this is not the only factor for the use of 'water' as a *vernacular* kind term. We here see a risk of mixing up our starting points again: we could very well have decided that the vernacular kind *water* did not have a chemical essence, although we could not at that point have decided that water did not *consist* of a chemical compound, or that H_2O is not a chemical kind.

LaPorte takes an *ex ante* perspective, looking from the conceptual situation and the knowledge existing in 1750. The eventual outcome of the decision was not available in 1750, even if the considerations from the point of chemistry might have been; the set of facts which would have a bearing on the decision were neither fixed nor weighted. At best, one could have a good hypothesis, an educated guess, as to the future result. But we can also take the *ex post* perspective from a position when a story can be told about the factors that (we believe) led to the outcome. Events were caused by certain factors and the remaining uncertainty is epistemological. Looking forward from $time_{Former}$, there was a decision to be taken about changing a concept one way or another in the light of new data; looking back from $time_{Later}$, we can see the path leading to the updated product.[508]

[507] See §6.2.

[508] Michael Jubien [2009], p.192, is therefore right in spirit, but not exactly correct in his formulation, when he writes: "What we could not discover is that it's essential to gold that it's an element and that it has atomic number 79." In my view, we *could* discover that gold has atomic number 79 *given* that it is an element in the sense of modern chemistry.

Åsa Wikforss raises another, related issue in her article "Are Natural Kind Terms Special?"[509] where she thinks that if we look closer at the thought experiments, we have to choose between a Putnamian view that an external[510] component contributes to the meaning and the view that there is a genuine, non-determined decision as to which way to go. She says that in a situation where the observational properties are identical, the external properties must determine the outcome – or be irrelevant.

> After all, to claim that we *could* go either way in Twin Earth scenarios is precisely to deny that the external feature plays a meaning-determining role: When the associated descriptions are the same…the external feature will be decisive, or else its role is null… [T]o support externalism it would have to be held that there was *no room for a decision*, that the underlying (external) essences were decisive.[511]

The issue is different, but the solution the same. I suggest that rather than saying that we have to choose between these two alternatives, we should say that the choice depends on our standpoint in terms of knowledge and concepts. If we use LaPorte's variation of the Twin Earth experiment,[512] we have the following situations (I will use the situation numbers as indices to the meaning of 'water' in the respective situations):

- Situation 0: we know that water on Earth is H_2O (essentially), but not anything about Twin Earth.
- Situation 1: we have discovered Twin Earth and its dominating liquid, but not yet its differing microstructure; it is assumed to be water.
- Situation 2: we know that what they call 'water' on Twin Earth has another microstructure than H_2O, namely D_2O.
- Situation 3: it is clear that what is called 'water' on Twin Earth is a kind of water.

[509] Wikforss [2010].
[510] "External" here means "non-mental".
[511] Wikforss [2010], p.77.
[512] See §6.2.

Situation 3 is then *either* a direct, determined consequence of Situation 2, *or* the result of a decision, Wikforss argues. But I think we can have both. In Situation 3, we know the reasons and motivations that led to the conclusion that deuterium is (a special type of) water, and it is not unnatural to talk about them as determining this outcome. In Situation 2, this might not be clear at all, and there is nothing in the term '$water_2$' that can help. We do not need to choose between the determination and the decision-making, we just need to index our argument with the right point of view. In Situation 2, a decision seems to be needed, as the D_2O case had not yet been put to the test. In LaPorte's terminology, 'water' ('$water_2$') is at this point vague as to whether D_2O is a variety of water or not. In Situation 2, it is *possible* that D_2O is not water. In Situation 3, on the other hand, the inclusion as a kind of (non-typical) water in the extension of '$water_3$' is determined by the reasons that led to the decision, in which the external factors, the chemical structures, were highly relevant. In Situation 3, it is *not* possible that D_2O is not water.

7.9. Conclusions

The commitments implicit in scientific activities serve to support the necessity-based argument for extra-theoretical, conceptual continuity. Natural kind terms are rigid, and scientific identifications necessarily true, because this is what scientists want to express, with their meta-paradigm commitments; and the methods of sciences justify our fallible belief in their statements.

Natural kind terms can be rigid in the same sense as proper names, because they function as names for abstract objects: the natural kinds. This is so, irrespective of what people living before and after us think or believe, and irrespective of the redubbings that take place at paradigm shifts.

Applying my framework, we see how oxygen can have existed in antiquity yet be a product of the Chemical Revolution. The framework shows who can be radically wrong about the meaning of a term (the lay user) and who cannot (the baptiser). It also combines a key role for decision-making with a determined continuity, as the continuity follows decisions taken for good, case-specific reasons, described in *ex post* stories.

I will conclude this thesis with a Coda, where I take one step further and apply the framework to illuminate a classic philosophical mystery, the Mind-Body problem.

8 Coda: Kripke's Critique of Physicalism

8.1 Introduction

In this last chapter I would like now to demonstrate the value of the framework I have developed in the previous chapters by showing how it clarifies a familiar topic in philosophy: the debate between Kripke and his opponents about Physicalism.

Towards the end of *N&N*, Kripke addresses the relationship between the body and the mind, and claims that his analyses "tell heavily"[513] against the usual forms of Physicalism. In "I&N" he states that physicalists are "up against a very stiff challenge"[514] in finding an explanatory model for their thesis, for there can be no discovery of a mind-body identity along the same model used during scientific discoveries. Some physicalists have responded that Kripke is misrepresenting them and assuming something physicalists have already rejected,[515] but the way Physicalism is formulated and defended today has been influenced by Kripke's criticism.

Daniel Stoljar suggests the following general definition of Physicalism:

> Physicalism is the thesis that everything is physical…Of course, physicalists don't deny that the world might contain many items that at first glance don't seem physical – items of a biological, or psychological, or moral, or social nature. But they insist nevertheless that at the end of the day such items are either physical or supervene on the physical.[516]

This definition leaves it open what 'physical' stands for. But in this chapter I will use it to mean 'described by physics', so that we can say that Physicalism holds that all entities in the world are

[513] Kripke [1981], p.155.
[514] Kripke [1971], p.163.
[515] Fred Feldman [1974], p. 676, writes regarding a particular premise that it is an "undefended, controversial premise that materialists have, and should have, rejected".
[516] Stoljar, Daniel [2017]. Initial paragraphs.

either postulated by physics or supervenes on such entities. Physicalism in this sense is a thesis about theories.[517]

Physicalism exists in many versions. I will here discuss two of these. The first, "The Identity Theory", claims that our mental state types are identical with events in the central nervous system ("brain events"). The second, "Functionalism", makes the same claim for functional states.

There is another important distinction between types of Physicalism. One version believes that Physicalism must explain our current set of vernacular concepts, sometimes called "folk psychology", as they are indispensable. Losing them, Jerry Fodor writes, would be no less than "the greatest intellectual catastrophe in the history of our species".[518] On the other hand Paul Churchland, defending the so-called "eliminativism", claims that

> our common-sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced…[519]

I will in this chapter present Kripke's argument, and how his premises and conclusions have been questioned by other writers. I will argue that a modest adjustment to the meaning of sensation terms puts Physicalism in a stronger position to respond to Kripke regarding for example 'pain'. However, this manoeuvre leaves a residue in terms of pain-experiences, where Kripke's arguments still apply.

When doing this, I will draw on the framework, arguments and conclusions I have outlined in my previous chapters to put up a final objection against Kripke; however, this will not be a defence of Physicalism as it exists today. I claim that the framework nonetheless gives us cause to hope that a solution one day might be found.

---

[517] I will *not* have in mind what can be called "materialism", by which I mean a thesis about what types of *objects* exist.
[518] Fodor [1987], p. xii.
[519] Churchland [1981], p.67.

8.2 Kripke and the (Psycho-Physical Type-Type) Identity Theory

Kripke's criticism of Physicalism comes at the end of the third lecture of *N&N* and is stated with the machinery he puts forward earlier in the book. It is formulated as a criticism against the Identity Theory, which was introduced at the end of the 1950s with pioneering articles by U T Place, Herbert Feigl and J J C Smart.[520] Most identity theorists want to explain the relationship between the mental states and brain events in terms of scientific identifications.

> When I say that a sensation is a brain process or that lightning is an electric discharge, I am using 'is' in the sense of strict identity. (Just as in the – in this case necessary – proposition '7 is identical with the smallest prime number greater than 5.'[521]

The proposed identity is on a *type* level, so every token/occurrence of pain is identical with a token/occurrence of (to use the standard example) c-fibre stimulation, and vice versa.[522]

Early formulations typically expressed the relationship as a *contingent identity*, but after "I&N" and *N&N* it has been commonly accepted that there is no such thing, and that all identity statements are *necessarily* true, if true at all. However, this has not been viewed as a serious problem for the Identity Theory, since what the earlier physicalists tried to achieve with contingent identities can, it is assumed, be expressed with the help of Kripke's necessary *a posteriori* instead.

Recall that Kripke sees scientific identifications as true identity statements, and natural kind terms as importantly similar to proper names. I have defended this view, with some qualifications. In particular, both types of terms are rigid, and for both, reference-determination can be separated from reference-fixing.

The following points therefore apply to both proper names and natural kind terms:

[520] Place [1956], Feigl [1958] and Smart [1959].
[521] Smart [1959], p.145.
[522] The designator 'c-fibre' is traditionally used in philosophy of mind and carries no exact empirical claim.

a. Rigid designators are terms that have the same reference in all possible worlds where they refer at all. Non-rigid designators do not.[523]
b. A statement is necessarily true if it is true in all possible worlds.
c. A statement is contingently true if it is actually true but there are possible worlds where it is not true.
d. (from **a** and **b**) Identity statements between rigid designators are necessarily true if true at all.[524]
e. (from **a** and **c**) Identity statements involving at least one non-rigid designator are not guaranteed to be necessarily true, even if they are true.
f. An object has its essential properties in all possible worlds. It has its contingent properties in some but not all possible worlds.

Consider now a typical scientific identification, such as:

S.$_{\text{Heat}}$ Heat is identical with molecular motion.

Kripke's primary target in his criticism against Physicalism is the Identity Theory that states:

I.$_{\text{Pain}}$ Pain is identical with c-fibre stimulation [in the same sense as heat is identical with molecular motion].

Kripke claims that the proposed mind-body identifications are importantly different from scientific identifications, that is, that Thesis I.$_{\text{Pain}}$ is importantly different from Thesis S.$_{\text{Heat}}$. He offers three different formulations of his argument, although they are not distinct: one using the notion of qualitative equivalence and the second rigidity; the third is expressed in terms of God's workload during the first week. I will refer to the first two in my semi-structured summary of Kripke's argument as it appears in the third Lecture of *N&N*:

1. (from Descartes) Everything that we can imagine without

[523] Kripke [1981], p.48.
[524] Kripke [1981], p.3.

contradiction is possible.[525]

2.I.$_{Pain}$ (from conceptual analysis) We can imagine pain existing without c-fibre stimulation, and vice versa.

3.I.$_{Pain}$ (from 1 and 2.I.$_{Pain}$) *It is possible that pain is not identical to c-fibre stimulation.*

But if we accept (1) unconditionally, the situation for the identity theorist is worse than than 3.I.$_{Pain}$ suggests, because:

4.I.$_{Pain}$ (from conceptual analysis) 'Pain' and 'c-fibre stimulation' are both rigid designators.[526]

5.I.$_{Pain}$ (from 3.I.$_{Pain}$ and **b**) Thesis I.$_{Pain}$ is not necessarily true.

6.I.$_{Pain}$ (from 5.I.$_{Pain}$ and **d**) *Pain is not identical to c-fibre stimulation.*

A problem arises immediately, however, because the argument above would also undermine scientific identifications such as 'heat is molecular motion'.

2.S.$_{Heat}$ (from conceptual analysis) We can imagine heat existing without molecular motion, and vice versa.[527]

3.S.$_{Heat}$ (from 1 and 2) *It is possible that heat is not identical to molecular motion.*

4.S.$_{Heat}$ (from conceptual analysis) 'Heat' and 'molecular motion' are both rigid designators.[528]

5.S.$_{Heat}$ (from 3.S.$_{Heat}$ and **b**) Thesis S.$_{Heat}$ is not necessarily true.

6.S.$_{Heat}$ (from 5.S.$_{Heat}$ and **d**) *Heat is not identical to molecular motion.*

That is certainly not what Kripke means.[529] He therefore adds two important qualifications to (1):

---

[525] Supported by Hume, who wrote: *[N]othing we imagine is absolutely impossible*." Hume [1968], p.32.

[526] Kripke [1981], pp.148-149: "[I]f something is a pain it is essentially so, and it seems absurd to suppose that pain could have been some phenomenon other than the one it is. The same holds for the term 'C-fiber stimulation', provided that 'C-fibers' is a rigid designator, as I will suppose here."

[527] Kripke [1981], p.99: "[C]haracteristic theoretical identifications, like 'heat is the motion of molecules', are not contingent truths but necessary truths, and here of course I don't mean just physically necessary, but necessary in the highest degree".

[528] Kripke [1981], p.136: "'Heat', like 'gold', is a rigid designator, whose reference is fixed by its 'definition'."

[529] Kripke [1981], p.99: "[C]haracteristic theoretical identifications, like 'heat is the motion of molecules', are not contingent truths but necessary truths, and here of course I don't mean just physically necessary, but necessary in the highest degree".

7. We cannot imagine what is in fact impossible.
8. As a special case of (7), we cannot imagine an object or event without its essential properties, since it is impossible for an object to exist without its essential properties.

These constraints, Kripke argues, separate scientific identifications (such as Thesis $S._{Heat}$) from the claims made by the Identity Theory (expressed in Thesis $I._{Pain}$). I will first present the rigidity-based version of his argument:

9. (from conceptual analysis) 'Heat' and 'molecular motion' are both rigid designators.
10. (from $4._{Heat}$ and **d**) The scientific identification 'Heat = molecular motion' is necessarily true, if it is true at all.
11. (from physics) 'Heat = molecular motion' is true.
12. (from 10 and 11) 'Heat = molecular motion' is necessarily true.
13. (from 7 and 12) We cannot imagine that heat is not molecular motion.

(13) notwithstanding, it is still true that I can *believe* that I imagine that heat is something else than molecular motion. There is at least an appearance of contingency. But Kripke can explain this using his distinction between reference-fixing and reference-determination. We can know how to use the word 'heat' based on observational properties, Kripke says, without knowing anything about chemistry, and we can use these properties to successfully refer to heat. The properties are contingent; if there were no sentient beings, heat would still exist, because heat, science has established, *is* molecular motion. Descriptions of contingent properties cannot determine the reference, but they can fix it.

In a second version of the argument, Kripke utilises the notion of epistemic equivalence to make the same point. If we believe that we imagine that we have heat experience not caused by heat, we *actually* imagine being in a situation that is epistemologically identical to a situation where we experience heat, but where we refer to something other than heat, something with the contingent property of being experienced as heat by us. This explains the appearance of contingency regarding the identity of heat with molecular motion.

The explanation of the appearance of contingency explains how we can make mistakes about scientific identifications. We can believe that we imagine that heat is something else than molecular motion, even if this actually is impossible. But for a sensation such as pain, the situation is different, Kripke argues. We refer to pain via an *essential* property, a property that pain must have to be pain at all, namely the human pain-experience. We cannot be in a situation epistemologically identical to being in pain without actually being in pain. We cannot imagine that pain is not experienced as pain. This situation does not allow the type of mistake described for heat.

Interim conclusion: So far, Kripke has argued that the appearance of contingency for Thesis $I._{Pain}$ is unexplained. We need to explain this, but cannot rely on the model used for scientific identifications. As long as we do not know how to do that, we have reasons to doubt Thesis $I._{Pain}$ that we do not have to doubt Thesis $S._{Heat}$.

8.3 The Cartesian Premise

Having arrived at this conclusion, we could go back to the very beginning and just reject the original premise (1) ("Everything that we can imagine without contradiction is possible"), an option Kripke recognises. We saw that he offers a model that leaves (1) in place as a default, but excludes cases that are impossible.

Point (8) says that we cannot imagine a situation where an object lacks its essential properties. Since being molecular motion is the essential property of heat, we cannot truly imagine that heat would exist without molecular motion. If we think we do, we are imagining something else.

> Can we then imagine a possible world in which heat was not molecular motion? We can imagine, of course, having discovered that it was not. It seems to me that any case which someone will think of, which he thinks at first is a case in which heat – contrary to what is actually the case – would have been something other than molecular motion, would actually be a case in which some creatures with different nerve endings from ours inhabit

> this planet…and in which these creatures were sensitive to that something else, say light, in such a way that they felt the same thing that we feel when we feel heat.[530]

But why is that – why can we not make mistakes about an object without involving *another* object? Just as I in an actual situation can have mistaken beliefs about facts without having to be thinking about something else, it seems I can in a counterfactual situation erroneously believe something in fact impossible to be possible, without referring to something else that is actually or possibly existing. This is the position taken by Sydney Shoemaker.

> [O]ne wonders why the explanation in terms of epistemic possibility is not enough; why must there be a genuine metaphysical possibility that is mistaken for the possibility of the situation imagined or conceived of?[531]

If we accept that our intuitions about what is possible can be mistaken, it is a short step (but not one that Kripke takes) to regard those intuitions as *appearances of possibility*, as Stephen Yablo suggests.[532] Intuitions would, according to this view, reflect our educated opinions, but not provide infallible knowledge. Kripke's position has the advantage of accounting for how the mistake is made – we mix up two qualitatively identical sets of facts – but maybe appearances of possibility can also be produced by other means. [533]

Now, if we accepted Yablo's suggestion, this would eliminate the force of (1) ("Everything that we can imagine without contradiction is possible"). Would it also undermine Kripke's argument against Physicalism as outlined above? The view that premise (1) is required is what Yablo calls "Textbook Kripkeanism",[534] to which Kripke himself, as Yablo acknowledges, may not be fully committed. However, in his comments on this distinction, Shoemaker states that rejecting Textbook Kripkeanism

[530] Kripke [1981], pp.131-132.
[531] Shoemaker [2011], p.341.
[532] Yablo [1993].
[533] Tamar Szabó Gendler and John Hawthorne [2002], p.9, suggest a weaker and more plausible version of the Cartesian premise (1): "[T]hat something is conceivable is at least a good indicator that it is possible."
[534] Yablo [2000].

> seems to require rejecting the contrast Kripke draws between the seeming possibility of heat without molecular motion and the seeming possibility of pain without C-fiber stimulation…[535]

I disagree. Kripke's argument that the alleged identity between mental and physical processes is importantly different from scientific discoveries does not rest on the Cartesian premise (1) only.

Let us look again at ($4.I._{Pain}$).

> $4.I._{Pain}$ (from conceptual analysis) 'Pain' and 'c-fibre stimulation' are both rigid designators.

As we have seen, this directly leads to problems for the physicalist if we keep premise (1). Removing (1) weakens the argument – but it does not destroy it. We are no longer entitled to ($6.I._{Pain}$) ("Pain is not identical to c-fibre stimulation"). But the asymmetry between heat and pain remains. I will call this "Kripke's asymmetry". We still have not addressed the underlying difference Kripke points to, that we refer to heat via human heat-experience, a contingent property of heat, and to pain via an essential property, the human pain-experience.

With this difference comes the unexplained feeling of contingency in the identification between mental states and brain events, claimed in Thesis $I._{Pain}$, and the doubts about whether Physicalism can be formulated as a scientific identification. This is indeed Kripke's own conclusion.[536] But there have been many objections. I will argue that none is totally effective.

8.4 Objections to Kripke's Analysis

Firstly, it is possible to deny that we refer to pain via an essential property, the human pain-experience.[537] Michael Levin argues that the reference to pain is *not* fixed by pain-experiences, but by functional role. He bases this on his modification of Wittgenstein's argument against a

[535] Shoemaker [2011], p.341.
[536] Kripke [1981], p.150: "I want to argue that, at least, the [psycho-physical] case cannot be interpreted as analogous to that of scientific identification of the usual sort, as exemplified by the identity of heat and molecular motion."
[537] See note 3, this chapter.

private language, which roughly says that the mental concepts cannot have been learned by links to inner experiences but must have been learned by links to public behaviour – their functional role.[538] Assuming that Wittgenstein's argument is valid, and assuming also that the functional role is a *contingent* property, we seem to have found a model that is closer to scientific identifications.

However, there are differences remaining between scientific identifications and psycho-physical identifications, even after Levin's analysis. We might have learned the concept 'pain' via its function and held a slot in that concept for the pain experience that typically accompanies the function. But this does not threaten Kripke's logic. The story can be described in Kripkean terms: initially the sample of pain has its reference fixed with the help of the functional role, but then adjusted to add Putnam's super-Spartans (pain but no behaviour)[539] and remove play-acting (behaviour but no pain). The original properties following from the functional role are eventually modified when we have realised that the essence of pain is the conscious pain experience.

A second physicalist response is to say that the problem is with *us*, not with the theory. David Papineau argues that there is nothing wrong with Physicalism as such.[540] What it needs is just a common acceptance.[541] I will discuss this response in §8.7.

A third option is to accept Kripke's statement that we cannot imagine that pain is not experienced as pain, but *still* claim that "Pain is identical with c-fibre stimulation", as the first part of Thesis $I._{Pain}$ says. If we give up premise (1), we could conceive a situation where pain-experience is an essential property, but where being c-fibre stimulation is *another* essential property. This view has been proposed by Thomas Nagel.[542] But his suggestion does not solve the problem with the lack of parallelism between the psycho-physical case and scientific

[538] Wittgenstein [1968]. Levin's version of Wittgenstein's §580 is Levin [1975], p.163, "[A]n inner process… stands in need of an outward reference-fixing".
[539] In his 1963 article "Brains and Behavior", included as chapter 16 in Putnam [1975], Putnam describes a society of "super-spartans" who feel and dislike pain, but have learned to suppress all responses to it for ideological reasons.
[540] Papineau [2008].
[541] He would also need to reject premise (1) ("Everything that we can imagine without contradiction is possible").
[542] Nagel [1979] and Nagel [2012].

identifications. This is an issue for the second part of Thesis $I._{Pain}$ ("*in the same sense* as heat is identical with molecular motion"), as pain would have a type of essential property that heat does not have. It is indeed doubtful whether this line would be consistent with Physicalism at all; Nagel himself has at different times expressed sympathies with both dual-attribute theories and with neutral monism, without fully supporting either of them.

A final possibility is to reject $4.I._{Pain}$ ("'Pain' and 'c-fibre stimulation' are both rigid designators"), as has been done by some functionalists. I will discuss this in the next section.

8.5 From the Identity Theory to Functionalism

Conceptually, a pure Identity Theory gets into trouble almost immediately. Just stating that mental states are brain events leave us with the problem to explain how the properties of a mental state are identical with the specifics of the brain event. We need to be convinced that mental states really have physical properties.[543] Furthermore, we need to be shown that the properties of brain events are sufficient to reduce and explain mental states, without any residue of mental properties, otherwise we have replaced a dualism of substances with a dualism of properties.[544]

In "Sensations and Brain Processes", Smart addresses these types of objection by suggesting an initial step of conceptual analysis, where descriptions of mental states are translated into a "quasi-logical or topic-neutral"[545] language, before any attempt to specify identities is made. This is already going beyond pure Identity Theory, heading in the direction of Functionalism.[546]

In Chapter 4, I described the process of a pre-scientific vernacular term turning into scientific one as a process where the meaning of term is adjusted before the scientific identification can

[543] See Jerome Shaffer [1961].
[544] Paul Feyerabend [1970], p.140, writes about Thesis $I._{Pain}$: "But this hypothesis backfires. It not only implies, as it is intended to imply, that mental events have physical features; it also seems to imply…that some physical events…have non-physical features."
[545] Smart [1959], p.150.
[546] I am here following McGinn's [2004] reading.

occur. This is the approach of (one version of) Functionalism, which suggests an initial step of conceptual analysis.[547] Smart's version of the Identity Theory has a very similar first step.

There is another point that is often mentioned as a major difference between the two types of physicalist theories. We saw that Kripke asks, rhetorically: "Can any case of essence be more obvious than the fact that *being a pain* is a necessary property of each pain?"[548] Nevertheless, this is in a sense what Functionalism denies.[549] I already quoted Lewis, who writes: "Pain might not have been pain...Something that is not pain might have been pain."[550] This is consistent with Functionalism, which denies that pain is identical with brain events on a type level. It does not, however, rule out that a pain token in a human being is identical to a brain event token (for example c-fibre stimulation). In line with MRT, the point is that there could in principle be pain tokens identical with tokens of *other* types of physical realisation, in an alien, or in a computer, perhaps. This is because for Functionalism, in contrast to Identity Theory, the c-fibre identity is not *necessary* for being in pain (although it may be *sufficient*). Functionalism does not rule out non-material implementations; it has no view on this. But it is compatible with materialism.[551]

In "Reduction of Mind", Lewis goes on to argue: "Kripke…vigorously intuits that some names for mental states, in particular 'pain', are rigid designators…I myself intuit no such thing".[552] Lewis here seems to believe he is making a similar point as we just made above about the token, but not type, identity of pain with c-fibre stimulation. But the way Lewis puts the point is in fact stronger, as he is rejecting ($4.I._{Pain}$) ("'Pain' and 'c-fibre stimulation' are both rigid designators"). If we accept Lewis's view, we have an argument against Kripke; but this comes at a high cost.

[547] This includes the philosophers defending the so-called "Canberra Plan".
[548] Kripke [1981], p.146.
[549] Kripke's critique is thus relevant to Functionalism as well, though he only mentions the theory. Kripke [1981], p .45, note 74: "Another view I will not discuss, although I have little tendency to accept it…is the so-called functional-state view of psychological concepts."
[550] Lewis [1980], p.125
[551] Ned Block [2015] criticises Functionalism for not being a proper Physicalist theory, because it does not reduce mental states to physical events.
[552] Lewis [1994], p.418-419.

Rejecting rigidity risks leaving Functionalism without an account of pain on a type level, and this is not doing justice to the theory.[553]

Functionalism, I believe, is better construed as holding that 'pain' rigidly refers to a kind whose essence is not microstructural. The MRT is naturally combined with pluralism in essence types. Using my terminology, Functionalism claims that mental states have a *functional* essence type. We see an example in Lewis's definition of 'pain':

> The concept of pain...is the concept of a state that occupies a certain causal role...a state apt for being caused in a certain way by stimuli plus other mental states and apt for combining with other mental states to jointly cause certain behavior.[554]

If we call a state of the type Lewis mentions in this quote, a "functional state", my interpretation of Functionalism says that the identity statement it puts forward features a rigid term that designates a (type of) functional state rather than a (type of) brain state, replacing Thesis $I._{Pain}$. with Thesis $F._{Pain}$:

$F._{Pain}$ Pain is identical with a functional state, call it "functional state F".

We need the corresponding update of ($4.I._{Pain}$), which becomes:

$4.F._{Pain}$ 'Pain' and 'functional state F' are both rigid designators.

If the essential properties of mental states are functions,[555] the Identity Theory is barking up the wrong tree, essence-wise.

Can Functionalism handle Kripke's criticism? I said earlier that denying that 'pain' is a rigid designator *would* give us a counter-argument against Kripke, but that it risks leaving Functionalism without an account of mental state types. In my interpretation Functionalism does

[553] I here take it for granted that a theory of the mind should offer explanatory power at a type level; see Jaegwon Kim [1998], p .7, for arguments. Lewis [1980] compromises and make definitions of mental states species specific.
[554] Lewis [1980], p.288.
[555] That is, if a theory where mental states are identified in terms of functional states offers more powerful explanations.

provide such an account, according to which 'pain' rigidly refers to a natural kind with functional essence, rather than one with a microstructural essence. As a result, I will argue in §8.7, Functionalism is open to MRT-inspired criticisms, just as the Identity Theory is. Ultimately, I will suggest, Functionalism and the Identity Theory are in the same position with respect to Kripke's asymmetry. I will therefore in the rest of this chapter usually talk about "Physicalism" in general, and not differentiate between the Identity Theory and Functionalism.

In the next section I discuss an idea that addresses Kripke's asymmetry, and which, I believe, is the best move available for Physicalism of either kind.

8.6 Separating Pains from Pain-Experiences

I claim that physicalists are not just proposing a theory change; they are also, and more radically, proposing one that includes changes to the meanings of terms. The meaning of the natural kind term 'pain' described by Lewis is similar but *not identical* to our current natural kind term 'pain'. To illustrate, we can look at William Lycan's way of elaborating the functionalist's case.

Lycan distinguishes "my *impression* or awareness that I am in pain, or my occurrent belief that I am in pain, from pain itself".[556] He continues: "It might be objected, by one who champions the incorrigibility and 'transparency' of my beliefs about my own mental states, that the pain and the awareness necessarily co-occur, and therefore are one and the same state"[557] – but finds this "far from obvious".[558] Referring to David Armstrong's analysis,[559] Lycan adds: "[O]ne can conceivably have pains of which one is unaware, and…seem to be aware of a pain that one does not in fact have."[560]

The incorrigibility thesis that Lycan mentions *is* of course central to what Kripke assumes, and what causes the discrepancy between the pain case and the heat case. I would argue that the distinction proposed by Lycan signifies a change of the meaning of 'pain'. Kripke's asymmetry

[556] Lycan [1974], p.682.
[557] Lycan [1974], p.682.
[558] Lycan [1974], p.682.
[559] Armstrong [1993].
[560] Lycan [1974], p.683.

seems to boil down to the lack in our current language of a distinction between the sensation and the phenomenon experienced, and a corresponding intermediary between the state and our knowledge of the state. This creates a problematic difference between identity theses put forward by any version of Physicalism and scientific identifications, even if we do not accept premise (1) ("Everything that we can imagine without contradiction is possible") without modifications. But this difference appears easy to fix. A distinction between pain and the experience of pain is neither dramatic nor counter-intuitive, and it would serve a good purpose.

It would not be unnatural to insert such a distinction to make the two cases, heat and pain, analogous. It is already correct English to say "I experience pain", as well as to say "I am in pain". If we accept that it has been established that pain is in fact functional state F, the need to distinguish this from the experience would remain; we would more frequently talk about experiences of pain in an analogous way to experiences of heat. With this distinction, we can imagine a situation where a creature is in pain without having pain sensations, and a situation where it has pain sensations without being in pain.

The physicalist can make use of Lycan's distinction to avoid Lewis's paradoxical expression "Pain might not have been pain"; we can then reformulate Lewis to state the logically impeccable: "A pain token might not have caused a pain-experience...something that is a pain-experience might not have been caused by a pain token."

A possible criticism of this move could be that it avoids Kripke's argument against Physicalism by begging the question, since it is changing the meaning of terms, rather than sticking to "our language". But if Kuhn is right, this is exactly what tends to happen in theory changes, rather than an *ad hoc* manoeuvre to avoid a critical argument. If the new theory has superior explanatory power, the move is defendable.

One school of thought looks at conscious states as perceptions.[561] Applied to sensations, this idea has been thought to run into problems with an asymmetry that is due to the essential property of

[561] See Locke [1975] and Armstrong [1993].

pain being felt as pain.[562] However, if we separate pain from pain-experience, the issue does not arise. Knowledge of these events would depend on empirical information acquired by an individual at a time. It would be fallible: we can have pain experiences without pains, and there can be pains of which we are not aware.[563] Separating pain from pain-experience, we make pain open to objective, scientific investigation, and remove a problem from a physicalist theory. Pain-experience, not pain, is perceptual. So construing pain-experiences (and other sensation-experiences) as perceptions is certainly a logical possibility, but I will not pursue this further here.

Rather, I return to a worry. Even if it makes sense to group pain-experiences and other sensations with perceptions,[564] this does not mean that our problem has gone away; Kripke's asymmetry is still there. As I said, the distinction between pain and pain-experience removes an obstacle to Physicalism that made its theses unexplainably different from scientific identifications. But Kripke's asymmetry is still relevant and threatening: no longer against *pain*, but against *pain-experience*. We have to consider the nature of these experiences further.

8.7 The Nature of Pain-Experiences

The issue with pain-experiences is not on a *token* level; it seems natural to regard those as a product of a basic biological function, not in need of much analysis. But Nagel points out that there is a *type*-level issue. There are generalisations connected with pain-experiences, which we can understand only if we have had a similar experience ourselves.

> It is often possible to take up a point of view other than one's own, so the comprehension of such facts is not limited to one's own case. There is a sense in which phenomenological facts are perfectly objective: one person can know or say of another what the quality of the other's experience is. They are subjective, however, in the sense

[562] See Murat Aydede [2009] for a discussion of this idea.
[563] This implies that what is perceived is *pain* rather than for example tissue damage.
[564] Aydede [2009], §3.2.: "There may be philosophical problems about how privacy, subjectivity and incorrigibility are possible in a completely physical world, but if there are such problems, they are general problems about having perceptual experience of any kind, not necessarily pertaining to pains and other intransitive bodily sensations."

> that even this objective ascription of experience is possible only for someone sufficiently similar to the object of ascription to be able to adopt his point of view – to understand the ascription in the first person as well as in the third, so to speak.[565]

It is not that we cannot say *anything* about pain-experiences; we can describe and compare them. It is just that this will not be fully comprehensible for someone who has not felt in a similar way herself. The article I just quoted is called "What Is It Like to Be a Bat?". With this example, Nagel wants to show the insurmountable problems for us to understand the conscious experience of a radically different creature. Bats undoubtably have conscious states, but these are alien to us.

Colin McGinn describes the same feature in terms of Russell's distinction between knowledge by acquaintance and knowledge by description.[566] Neither implies the other. If a concept is acquaintance-based, then we do not fully grasp it unless we have personal experience of what it refers to. All terms referring to conscious mental states belong to this category, which also includes other terms, such as those for colours (this is Russell's example); a person born blind, presumably, cannot fully understand the meaning of 'red'. We could say that the stereotype of 'pain' in the sense of 'pain-experience' is only fully mastered if we have knowledge by acquaintance of such experiences. We are not completely linguistically competent with the term if we lack them. We can extrapolate from our own case to creatures similar to us, but the more different they are, the less adequate this method is.

The point of the bat example is exactly that; we are not equipped to understand what it is like to be a bat, because we do not have knowledge by acquaintance of bat experiences and could therefore not competently use terms that describe them – which does not mean that we should deny the existence of these experiences. When I try to imagine being a bat, Nagel says: "I am restricted to the resources of my own mind, and those resources are inadequate to the task."[567] He adds:

---

[565] Nagel [1974], pp.441-442.
[566] "Knowledge by acquaintance" is Russell's name for our special access method to mental states, by which states of the phenomenological essence types access their objects. See Russell [1910-1911].
[567] Nagel [1974], p.439.

> It would be fine if someone were to develop concepts and a theory that enabled us to think about those things; but such an understanding may be permanently denied to us by the limits of our nature. And to deny the reality or logical significance of what we can never describe or understand is the crudest form of cognitive dissonance.[568,569]

Not to be guilty of the cognitive dissonance Nagel talks about, we should accept that there are in fact pain-experiences there to be explained. But Nagel and McGinn *indicate* the existence of pain-experiences, they do not *explain* them. This is the crucial point: our lack of tools, of a theoretical framework to analyse and define pain-experiences.

In Chapter 7, I interpreted Kripke as saying that natural kind terms like 'water' had a slot to be filled before H_2O was found to be its essence. In my terms, the Chemical Revolution had tentatively identified the essence type, subsequently confirmed by the discovery of a chemical essence, which also led to a change of the meaning of 'water'. In arguing against the Identity Theory, Kripke is (on my interpretation) claiming that the difference between 'pain' and 'water' is that our current natural kind terms for sensation states *do not* have a slot that can be filled by a microstructural (in this case neurophysiological) essences.[570] The same argument can be applied against Functionalism and functional essences.

As McGinn puts the point: "A successful reduction of water to H_2O does not leave open the option that water and H_2O might yet be separate substances; it is not interpretable as merely stating a correlation. In the same way any adequate reduction of consciousness must make dualism a non-starter".[571] Joseph Levine[572] has coined the phrase "the explanatory gap" for this difference from states like 'pain'. But I would put it slightly differently than McGinn does, because a physicalist could just accept what McGinn says about reduction, while insisting that what is needed is a *change of meaning* for 'pain' that excludes pain-experience. The response has

[568] Nagel [1974], pp.440-441.
[569] McGinn [2004], p.51, believes that similar "limits of our nature" kick in already when we try to understand human mental states, and is therefore pessimistic about the success of future sciences in this area: "What I am suggesting, basically, is the existence of (humanly) unknowable conceptual connections between mind and brain."
[570] In Kripkean terms, there is no enterprise that includes such a slot.
[571] McGinn [2004], p.15.
[572] Levine [1983].

merits, because as I have earlier argued, the initial creation of a scientific abstract object such as water is not just a discovery, but involves an element of decision, and that some features that do not fit well with a new theory can be excluded.[573] But in this case, it would leave an important residue.

I earlier referred to Papineau's suggestion that there is nothing major wrong with Physicalism, what is needed is our acceptance, and I will now return to this point. Underlying our Cartesian intuition that it would be possible for the mind to exist without the brain and vice versa, he claims, sits the fact that we have not yet fully *accepted* the identity on a theoretical level. This is the case: the identity between water and H_2O is now part of the water stereotype, but physical identities with brain events are not (at least not yet) part of mental natural kind concepts. There might be understandable psychological reasons for this, but this does not necessarily mean that there is anything wrong with Physicalism as a theory. "[F]ully believing that pain is c-fibres firing will destroy any epistemological possibility of its being different, along with any metaphysical possibility thereof".[574] He adds: "Many things that strike human beings as intuitively false nevertheless turn out to be true."[575]

I have earlier, following Kuhn and LaPorte, found that decision-making plays a crucial role at certain stages of scientific progress, scientific revolutions. Could we not *decide* to accept Physicalism, in the same way as we have taken decisions to accept other scientific theories? If Papineau is right, this would take care of any explanatory gaps, as "the so-called 'explanatory gap' is simply a manifestation of an intuitive conviction that dualism is true".[576] Our resistance at this point could be a case of delayed acceptance, which Kuhn points out can follow the introduction of a new theory.

The difficulty with this suggestion is that the decisions in the history of science have been taken for good specific reasons, clear in *ex post* stories, where we see the explanatory power of theories

[573] See the description of 'acid' in §4.3.
[574] Papineau [2008], §6.4.
[575] Papineau [2008], §7.
[576] Papineau [2008], §1.

enhanced. It would certainly be rational to overcome whatever prejudice we might have and accept a theory of the human mind with greater explanatory power. But neither the Identity Theory nor Functionalism deliver in this respect; they both fail to take sensation experiences into account.

The problem for Physicalism is that pain experience is not just the belief that I am in pain, not just an *awareness* of this condition. The difference between a pain that I am not aware of and one that I *am* aware of is not a matter just of having access to data in an abstract, cognitive sense. If we take that route, we lose the actual *painfulness* in the analysis. It is not awareness of pain that is needed to complement the physical state; it is the experience of pain. It is not just a matter of having the information; it is also a matter of the suffering. Functional states do not seem useful for defining painful experiences. Indeed, the intuitive force of MRT against the Identity Theory is, regarding sensation-experiences, equally strong against Functionalism. It appears just as plausible that pain-experiences can supervene on multiple functional states as on multiple neurological ones.[577]

Scientific decision-making often involves meaning changes of key terms, but acceptance of such decisions is dependent on a continuity for good reasons, which is a problem in this case. We are naturally reluctant to accept a theory that does not have enough explanatory power to cover relevant, known data, data we know by acquaintance, to use Russell's term. Functional states, like brain events, fail this criterion for being essences of pain-experiences. Instead, pain-experiences have what we could call a "phenomenological" essence type.[578]

8.8 Physicalism, Pluralism and Our Language

So far, I have presented two versions of Physicalism, Kripke's criticism of Physicalism (in particular the Identity Theory) and some objections from the literature against Kripke's analysis. In the last two sections, I have tried to strengthen Physicalism with the help of a distinction

[577] Putnam make this argument in Putnam [1988]. See also Ned Block [1978].
[578] While their authors draw different conclusions, some much-discussed thought experiments want to illustrate this point, that conscious experiences cannot be functionally defined without residue. These include Block's [1978] Chinese Nation, Searle's [1980] Chinese Room, Jackson's [1986] Mary, and Chalmers' [1996] zombies.

borrowed from Lycan, but also suggested that a residue remains unexplained. I therefore concluded that neither neurophysiological nor functional states provide the right essence type for experiences such as pain-experience. In the following two sections, I will connect this Coda with the previous chapters, and apply my temporal indices, the analysis of conceptual change, and the conclusions about pluralism. This will put me in a position to recognise the strength of Kripke's argument while still holding out some hope for physicalists – by returning to Kuhn's sociological thesis. I will start with the indices and the objection against Kripke that they suggest.

We could accept Kripke's arguments, but question their scope. If my previous reasoning about temporal indexing of natural kind terms is sound, Kripke's conclusions above, even if they are right (as I believe they are), are bound by the scope of our paradigm and our current concepts. We immediately see this if we replace them with '$\text{pain}_{\text{Today}}$', '$\text{heat}_{\text{Today}}$', and so on.

If terms such as 'pain', used by Kripke in his argument against Physicalism, carry implicit temporal indices, Kripke's argument has a limited scope. Essences are affected by decision-making. We can say that mental states, as we define them, cannot be identified with anything posited by current neurophysiology or cognitive science. We can also say that it is unlikely that "more of the same" normal science (in Kuhn's sense) can address this. But this does not rule out that new paradigms that one day replace current ones will alter this situation – without changing the subject. The power of Kripke's thought experiments stays within the current paradigm. Therefore, we might conclude that Kripke's argument against Physicalism fails.

To do that, however, would be a mistake. I stand by my previous analysis, but there is nevertheless a flaw in this conclusion. In the beginning of this chapter I wrote: "Physicalism holds that all entities in the world are either postulated by physics or supervenes on such entities." If we by 'physics' mean a theory that builds further, in a normal-science mode, on the existing science with this name, the counter-argument against Kripke is undermined. Since 'Physicalism' refers to the current paradigm of physics, it is *just as paradigm-bound* as Kripke's criticism of it.

If, on the other hand, Physicalism is based on the hope that *future* paradigms will address current problems, the term 'Physicalism' is a misnomer, for it would *not* be based on physics. If we, contrary to what I assumed above, try to fix this by allowing 'physics' to mean 'a future scientific theory rich enough to explain all that we want explained in the philosophy of mind', but otherwise staying unspecified, Kripke's argument indeed fails, but only because the physicalist thesis becomes impossible to refute. We would have stipulated the existence of a theory that solves our problem, but of which we have no independent knowledge. McGinn defines this future theory as

> the doctrine that the mental is reducible to what would feature in an ideal theory of the world…But that is totally vacuous: *of course* the mental is so reducible, even if mental terms themselves figure in that ultimate theory. That final theory might invoke very different kinds of entities and principles from those we speak of today…[579]

At this point it is useful, I suggest, to recall Kuhn's sociological thesis, where I started off in Chapter 1. Scientists are working within a paradigm during normal-science periods, and paradigms define the meanings of key terms, but these periods are followed by crises and revolutions. Essences can change with new theories, and properties that are essential today can become contingent in the future, as I argued in Chapter 4. We have no notion of what a future theory of the human mind will look like, but that is typical before a paradigm shift, during a crisis phase. The logic that leads to the eventual result is only clear *ex post*.

It might be possible and useful for a future revolutionary science to change the meanings of natural kind terms without abandoning them. But the latter option is also open. There can be different kinds of conceptual changes caused by changes in scientific theories,[580] including these:

- Sometimes terms in the old theory are retired as no longer needed, without being empirically refuted.
- Sometimes terms keep some part of their meaning, while another is given up.

[579] McGinn [2004], p.17.
[580] LaPorte [2004] gives examples of different types.

- Sometimes conceptual changes imply modal and epistemological changes for statements including these terms.

Future theories will no doubt imply conceptual changes for current terms and include new terms. If a radically new theory comes along, which covers the data and offers explanatory power for a set of mental states similar to our current classification, acceptance would probably be forthcoming. Would another classification be more powerful, we might get a different result. We cannot know the outcome of future decisions.

We should also ask why many writers feel that there *must* somehow be a theory that explains pain and pain-experience in terms of physics, without residue. Putting the question this way suggests another appeal to the PTT. The view implied by this notion is that science should aim to resemble the theoretically possible perfect theory, which is integrated and all-encompassing, and whose postulates corresponds one-to-one with the objectively existing entities in nature. Anything left out, not possible to integrate and reduce, is according to the PTT a weakness to be corrected. There is a PTT echo in the McGinn quote above, when he speaks about an "ideal" or a "final" theory. McGinn believes in the perfect theory for mental states, but he combines this belief with a strong pessimism about our ability to make progress in this direction.

The previous chapters have suggested another view of science, consisting of many different approaches and projects, where postulates are chosen for their explanatory power. This approach allows different, approximately true descriptions of reality, improving over time, without insisting on total coverage and total integration. It is not threatened if it cannot exhaust first-person conscious experiences with physics (current or future), as long as we have a powerful theory that can explain them in its own way. So if we instead look from a standpoint of pluralism, for which we found good arguments in Chapter 5, we should say that a future theory addressing human mental states may offer reduction and integration; but then again, it might not. It is possible that there instead will be multiple, unintegrated theories, successfully addressing and explaining different aspects of the human mind, for different reasons.

8.9 Conclusions

I have in this chapter applied the framework outlined in previous chapters to the Mind-Body problem and I will now summarise the result.

Accepting previously the validity of thought experiments and the extension of Kripke's machinery to natural kind terms, I also accept the semantic conclusions Kripke and Putnam draw: natural kind terms are rigid and scientific identifications necessarily true, if true at all. This must also be the case for scientific identifications that feature mental states. But Kripke's critique of Physicalism shows that for sensation terms, there is an asymmetry that needs to be explained.

We can regard the difference between the Identity Theory and Functionalism as the difference of which essence type is allocated to mental sensations: a microstructural or a functional. Both the Identity Theory and Functionalism arrive at these essence types via a two-step approach where the initial step consists of conceptual analysis. But this approach, from a vernacular starting point, shows weaknesses. The vernacular sensation terms, with phenomenological essences, are not open for scientific identifications with physical events. There is a need for a scientific starting point, with a redubbing of the relevant natural kinds, where natural kind terms take on a meaning given by a scientific theory. I pointed to a start of this exercise, separating sensations from the experience of them. But Kripke's asymmetry still remains for experience terms.

I have suggested that Kripke's claims about the asymmetry is paradigm-bound, consistently with my argument that natural kind terms, their rigidity and the necessity of scientific identifications are paradigm-bound.[581] This is the reason we can combine the Kripke-Putnam semantics with Kuhn's scientific phases and description of paradigm shifts. But this is of no help for the current-day physicalist, as his theory is equally paradigm-bound. Future progress is required and expected – scientific progress has not stopped – but Physicalism, if this term is to have any meaning, has to be judged on what it says today. It is possible that Kripke's objections will be met by a comprehensive future theory about mental states, producing appropriate scientific identifications, but that theory will not be $\text{Physicalism}_{\text{Today}}$, based on $\text{physics}_{\text{Today}}$. Due to

[581] §7.4.

closeness of our current sensation terms to scientific identifications, normal science, in Kuhn's sense, cannot deliver such identifications.

In other words, we are now in a Kuhnian crisis, where the available theories do not fit the established facts. The situation, as Kripke says, is "wide open and extremely confusing."[582]

Papineau points to the need for decision-making to put the transition to a new paradigm in motion, and we have seen that decisions can implement change as well as provide continuity. But these decisions are decisions taken for a good reason, to introduce a theory that increase the explanatory power for the purpose of enquiry. To get a satisfactory theory of the human mind we need revolutionary science with a radical paradigm shift. But such revolutions do occur. We cannot beforehand know which decisions will be taken in future paradigm shifts, and that also applies to terms for mental states. Conceptual changes happen at decision points where the relevant factors and their weighting are unknown before the event. The decisions will be taken for good reasons, which will be clear and justified *ex post*. If this is right, the discussion between eliminativists and their opponent is pre-mature. There is also no call for Fodor's worry, as a theory could only replace folk psychology terms by terms with greater explanatory value.

We cannot take for granted that the scientific revolution for mental state terms will come in the form of an integrated, comprehensive theory. We do not know what a future theory explaining mental states and experiences of mental states will look like, if it can be reduced to a future physical science, or if a reduction indeed would be what is deemed useful for future scientists. With the view of scientific endeavours as separate projects, trying to answer different questions, choosing tools according to explanatory power for their purposes, the view of a theory (current or future) that does not reduce to physics (current or future) is no longer a failure, and the potential non-existence of a grounding microstructure no longer a threat.

Pluralism teaches us that natural kinds are postulated given a purpose of enquiry, and that also theories without integration, indeed also mutually inconsistent theories, can deliver success and

[582] Kripke [1981], p.155, note 77.

capture approximate truths about the world. By increasing the explanatory power in their field of enquiry, new theories increase our knowledge, that is, they give us scientific progress.

Bibliography

Andreas, Holger (2017) "Theoretical Terms in Science", *Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2017/entries/theoretical-terms-science/. Accessed on 27 October 2019.

Aristotle *Metaphysics*, Loeb Classical Library. Vol. 1. ebook.

Armstrong, D.M. (1993*) A Materialist Theory of the Mind*, London : Routledge. First published 1968.

Armstrong, D.M (1997) "Against Ostrich Nominalism: A Reply to Michael Devitt," in D. H. Mellor and A. Oliver (eds.) (1997) *Properties*, pp. 101-111, Oxford : Oxford University Press. First published 1980.

Aydede, Murat (2019) "Pain", *Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2019/entries/pain/. Accessed on 27 October 2019.

Beebee, Helen and Nigel Sabbarton-Leary (eds.) (2010) *The Semantics and Metaphysics of Natural Kinds*, New York ; London : Routledge.

Berger, Alan (ed.) (2011) *Saul Kripke*, New York : Cambridge University Press.

Bird, Alexander (2004) "Kuhn on Reference and Essence", *Philosophia Scientiæ*, Vol. 8, No. 1, pp. 39-71.

Bird, Alexander (2010) "Discovering the Essences of Natural Kinds" in Beebee and Sabbarton-Leary (eds.) (2010), pp. 125-136.

Bird, Alexander and Emma Tobin (2018) "Natural Kinds", *Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/natural-kinds/. Accessed on 27 October 2019.

Block, Ned and Jerry Fodor (1972) "What Psychological States Are Not", *Philosophical Review*, Vol. 81, No. 2 (April), pp. 159-81.

Block, Ned (1978) "Troubles with Functionalism", *Minnesota Studies in the Philosophy of Science*, Vol. 9, pp. 261-325.

Block, Ned (2015) "The Canberra Plan Neglects Ground" in Terence Horgan, Marcelo Sabatés & David Sosa (eds.), *Qualia and Mental Causation in a Physical World:*

Themes from the Philosophy of Jaegwon Kim, pp. 105-133, Cambridge : Cambridge University Press.

Boyd, Richard (1999) "Homeostasis, Species, and Higher Taxa" in Wilson, Robert (ed.) (1999) *Species: New Interdisciplinary Essays*, pp. 141-185, Cambridge, Massachusetts : MIT Press.

Brendel, Elke (2004) "Intuition Pumps and the Proper Use of Thought Experiments", *Dialectica*, Vol. 58, No. 1, pp. 89–108.

Brown, James Robert and Yiftach Fehige (2019) "Thought Experiments", *Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2019/entries/thought-experiment. Accessed on 2 November 2019.

Bruce, Steve (1999) *Sociology, A Very Short Introduction*, Oxford : Oxford University Press.

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

Cartwright, Nancy (2004) "Causation: One Word, Many Things", *Philosophy of Science* Vol. 71, No. 5 (December).

Chakravartty, Anjan (2007) *A Metaphysics for Scientific Realism: Knowing the Unobservable*, Cambridge ; New York : Cambridge University Press,.

Chakravartty, Anjan (2017) "Scientific Realism", *Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2017/entries/scientific-realism. Accessed on 27 October 2019.

Chalmers, David (1997) *A Conscious Mind: In Search of a Fundamental Theory*, Oxford : Oxford University Press. Paperback. First published 1996.

Chang, Hasok (2014) *Is Water H_2O? Evidence, Realism and Pluralism*, Dordrecht : Springer.

Churchland, Patricia Smith (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, Cambridge, MA ; London : MIT Press.

Churchland, Paul M. (1981) "Eliminative Materialism and the Propositional Attitudes", *The Journal of Philosophy*, Vol. 78, No. 2 (Feb), pp. 67-90.

Churchland, Paul (1985) "Conceptual Progress and Word/World Relations: In Search of

the Essence of Natural Kinds.", *Canadian Journal of Philosophy*, Vol. 15, pp. 1-17.

Cohen, Jonathan and Craig Callender (2009) "A Better Best System Account of Lawhood", *Philosophical Studies*, 145, No. 1 (July), pp. 1-34.

Davidson, Donald (1973) "Radical Interpretation", *Dialectica*, 27, No. 3/4, pp. 314-328.

Davidson, Donald (1973-1974) "On the Very Idea of a Conceptual Scheme", *Proceedings and Addresses of the American Philosophical Association*, Vol. 47, pp. 5-20.

Davidson, Donald (1984) *Inquiries into Truth and Interpretation*, Oxford : Clarendon Press.

Dennett, Daniel Clement (1980) "The Milk of Human Intentionality*", Behavioral and Brain Sciences*, 3, pp. 428-430.

Dennett, Daniel Clement (2015) *Elbow Room: The Varieties of Free Will Worth Wanting*, Cambridge, MA ; London : MIT Press. First published 1984.

Devitt, Michael (1983) "Realism and the Renegade Putnam: A Critical Study of Meaning and the Moral Sciences", *Noûs*, Vol. 17, No. 2 (May), pp. 291-301.

Devitt, Michael (2008) "Resurrecting Biological Essentialism", *Philosophy of Science*, Vol. 75, No. 3 (July), pp. 344-382.

Donellan, Keith (1983) "Kripke and Putnam on Natural Kind Terms", in Carl Ginet and Sydney Shoemaker (eds.) *Knowledge and Mind: Philosophical Essays*, pp. 84-104, New York ; Oxford : Oxford University Press.

Douven, Igor (1999) "A Critique of Putnam's Principle of the Benefit of Doubt", *Logic Group Preprint Series*, volume 195, pp. 1-12, Utrecht University. https://dspace.library.uu.nl/handle/1874/26931. Accessed on 27 October 2019.

Dupré, John (1995) *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*, Cambridge, MA : Harvard University Press. Paperback edition. First published 1993.

Ellis, Brian (2001) *Scientific Essentialism*, Cambridge : Cambridge University Press.

Evans, Gareth (1973) "The Causal Theory of Names", *Proceedings of the Aristotelian Society*, Supplementary Volumes, Vol. 47, pp. 187–208.

Evans, Gareth (1982) *The Varieties of Reference*, Oxford University Press, Oxford: New York.

Feigl, Herbert (1958) "The 'Mental' and the 'Physical'", in H. Feigl, M. Scriven and G. Maxwell (eds.), *Concepts, Theories and the Mind-Body Problem* (Minnesota

Studies in the Philosophy of Science, Volume 2), pp. 370-497, Minneapolis : University of Minnesota Press.

Feldman, Fred (1974) "Kripke on the Identity Theory", *Journal of Philosophy*, vol. 71, No.18 (October), pp. 665-676.

Fernández Moreno, Luis (2016) "Putnam's View on Reference Change Is Different from That of Kripke's", *Organon F*, Vol. 23, No. 3. Central and Eastern European Online Library. https://www.ceeol.com/search/article-detail?id=693557. Accessed on 27 October 2019.

Feyerabend, Paul [1970] "Comment: 'Mental Events and the Brain'", in *The Mind-Brain Identity Theory* [1970] C.V. Borst (ed.), pp. 140-141. London ; Basingstoke : The MacMillan Press.

Fine, Kit (1994) "Essence and Modality: The Second Philosophical Perspectives Lecture", *Philosophical Perspectives*, Vol. 8, pp. 1-16.

Fodor, Jerry (1974) "Special Sciences: Or the Disunity of Science as a Working Hypothesis", *Synthese*, Vol. 28, No. 2, pp. 97-115.

Fodor, Jerry (1987) *Psychosemantics : The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA ; London : MIT Press.

Frege, Gottlob (1980) "On Sense and Reference", in *Translations from the Philosophical Writings of Gottlob Frege*, Peter Geach and Max Black (eds.), Oxford : Basil Blackwell.

Friedman, Michael (2009) "Einstein, Kant, and the Relativized *A Priori*", in Michel Bitbol et al. (eds.), *Constituting Objectivity: Transcendental Perspectives on Modern Science*, pp. 253-267, Springer Science + Business Media B.V.

Friedman, Michael (2010) "Remarks on the History of Science and the History of Philosophy", in Paul Horwich (ed.) (2010), pp. 37-54.

Gendler, Tamar Szabó and John Hawthorne (2002) *Conceivability and Possibility*, Oxford ; New York : Oxford University Press. Kindle edition.

Ghiselin, Michael (1987) "Species Concepts, Individuality, and Objectivity" *Biology and Philosophy*, Vol. 2, No. 2 (April), pp. 127-143.

Godfrey-Smith, Peter (2003) *Theory and Reality: An Introduction to the Philosophy of Science*, Chicago ; London : University of Chicago Press.

Godfrey-Smith, Peter (2014) *Philosophy of Biology*, Princeton, Woodstock ; Oxon : Princeton University Press.

Hacking, Ian (1975) *What Does Language Matter to Philosophy*, Cambridge : Cambridge University Press.

Hacking, Ian (1983) *Representing and Intervening*, Cambridge ; New York : Cambridge University Press.

Hacking, Ian (2007) "Putnam's Theory of Natural Kinds and Their Names is Not the Same as Kripke's", *Principia*, pp. 1-24.

Hacking, Ian (2010) "Working in a New World: The Taxonomic Solution" in Horwich (ed.) (2010), pp. 275-310.

Hacking, Ian (2012) "Introductory Essay", in Kuhn (2012), pp. vii-xxxvii.

Hendry, Robin Findlay (2010) "The Elements and Conceptual Change" in Beebee and Sabbarton-Leary (eds.) (2010), pp. 137-158.

Horwich, Paul (ed.) (2010) *World Changes: Thomas Kuhn and the Nature of Science*, Pittsburgh : First University of Pittsburgh Press. Paperback.

Hughes, Christopher (2004) *Kripke: Naming, Necessity and Identity*, Oxford ; New York : Oxford University Press.

Hume, David (1978) *Treatise of Human Nature*, Selby-Bigge/Nidditch, Oxford : Oxford University Press.

van Inwagen, Peter (1990) *Material Beings*, Ithaca ; London : Cornell University Press.

Jackson, Frank (1986) "What Mary Didn't Know", *Journal of Philosophy*, 83, No. 5 (May), pp. 291-295.

Jackson, Frank (1994) "Finding the Mind in the Natural World", in Roberto Casati, B. Smith & Stephen L. White (eds.), *Philosophy and Cognitive Sciences*, pp. 227-49, Vienna : Holder-Pichler-Tempsky.

Jackson, Frank (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford ; New York : Oxford University Press.

Jones, Roger (1991) "Realism about What?", *Philosophy of Science*, Vol. 58, No. 2 (June), pp. 185-202.

Jubien, Michael (2009) *Possibility*, Oxford : Clarendon Press.

Kant, Immanuel (1933) *Critique of Pure Reason*, Basingstoke ; London : MacMillan Publishers.

Khalidi, Muhammad Ali (2015) *Natural Categories and Human Kinds: Classification in the Natural and Social Sciences*, Cambridge: Cambridge University Press. Paperback. First published 2013.

Kim, Jaegwon (1998) *Mind in a Physical World : An Essay on the Mind-Body Problem and Mental Causation*, Cambridge, MA : MIT Press.

Kripke, Saul (1971) "Identity and Necessity" in Milton Karl Munitz (ed.) (1971), *Identity and Individuation*, pp. 135-164, New York : New York University Press.

Kripke, Saul (1981) *Naming and Necessity*, Malden, MA : Blackwell Publishing. Paperback. First published 1972. Extended version 1980.

Kripke, Saul (2011) *Philosophical Troubles*, Collected Papers, Vol. 1, Oxford ; New York : Oxford University Press. ebook.

Kroon, Frederick (1985) "Theoretical Terms and the Causal View of Reference", *Australasian Journal of Philosophy*, Vol. 63, No. 2, pp. 143-166.

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

Kuhn, Thomas (1990) "Dubbing and Redubbing: The Vulnerability of Rigid Designation", in C. Wade Savage (ed.) (1990) *Scientific Theories*, Minnesota Studies in Philosophy of Science, Vol. 14, pp. 298-318. Minneapolis : University of Minnesota Press.

Kuhn, Thomas (2002) *The Road Since Structure*, Chicago : University of Chicago Press. Paperback. First published 2000.

Kuhn, Thomas (2012) *The Structure of Scientific Revolutions*, Chicago : The University of Chicago Press. Fourth Edition. First published 1962.

Kuukkanen, Jouni-Matti (2010) "Kuhn on Essentialism and the Causal Theory of Reference", *Philosophy of Science*, Vol. 77, No. 4 (October), pp. 544-564.

LaPorte, Joseph (2004) *Natural Kinds and Conceptual Change*, Cambridge : Cambridge University Press.

LaPorte, Joseph (2010) "Theoretical Identity Statements, Their Truth, and Their Discovery" in Beebee and Sabbarton-Leary (eds.) (2010), pp. 104-124.

LaPorte, Joseph (2013) *Rigid Designation and Theoretical Identities*, Oxford: Oxford University Press.

Laudan, Larry (1981) "A Confutation of Convergent Realism", *Philosophy of Science*, Vol. 48, No. 1 (March), pp. 19-49.

Laudan, Larry (1984) "Explaining the Success of Science" in James T. Cushing, C.F. Delaney, Gary M. Gutting (1984) (eds.) *Science and Reality: Recent Work in the Philosophy of Science*, pp. 83–105. Notre Dame : University of Notre Dame Press.

Levin, Michael E. (1975) “Kripke's Argument Against the Identity Thesis”, *Journal of Philosophy*, 6 (March), pp. 149-167.

Levine, Joseph (1983) “Materialism and Qualia: The Explanatory Gap,” *Pacific Philosophical Quarterly*, 64, No. 4 (October), pp. 354-361.

Lewis, David (1971), “Counterparts of Persons and Their Bodies,” *Journal of Philosophy*, 68, No. 7 (April), pp. 203-211.

Lewis, David (1980) “Mad Pain and Martian Pain” in Ned Block (ed.) (1980) *Readings in Philosophy of Psychology*, Vol. I, pp. 216-222. Cambridge, MA : Harvard University Press.

Lewis, David (1986) *On The Plurality of Worlds*, Oxford ; Malden, MA : Blackwell.

Lewis, David (1994) “Reduction of Mind”, in Samuel Guttenplan (ed.), *Companion to the Philosophy of Mind*, pp. 412-431. Oxford : Blackwell.

Li, Chenyang (1993) “Natural Kinds: Direct Reference, Realism, and the Impossibility of Necessary a Posteriori Truth”, *Review of Metaphysics*, Volume 47, No. 2 (December), pp. 261-276.

Linsky, Bernard (2011) ”Kripke on Proper and General Names” in Alan Berger (ed.) (2011), pp. 17-48.

Locke, John (1975) *An Essay concerning Human Understanding*, P.H. Nidditch (ed.), Oxford : Oxford University Press. First published issue in 1689.

Lycan, William (1974) “Kripke and the Materialists”, *Journal of Philosophy*, Vol. 71, No.18 (October), pp. 677-678.

Mackie, Penelope (2006) *How Things Might Have Been: Individuals, Kinds, and Essential Properties*, Oxford : Clarendon Press. ebook.

Magnus, P.D. (2012) *Scientific Enquiry and Natural Kinds: From Planets to Mallards.* New York ; Basingstoke : Palgrave Macmillan. Kindle Edition.

Marcus, Ruth Barcan (1961) “Modalities and Intensional Languages”, *Synthese*, Vol 13, No. 4 (December), pp. 303-322.

Masterman, Margaret (1970) “The Nature of a Paradigm”, in Imre Lakatos and Alan Musgrave (eds.) (1970), *Criticism and the Growth of Knowledge*, pp. 59-89. London : Cambridge University Press.

McGinn, Colin (2004) *Consciousness and Its Objects*, Oxford : Clarendon Press. ebook.

McGinn, Colin (2015) *Philosophy of Language*, Cambridge, MA ; London : MIT Press.

Mill, John Stuart (1973-1974) *A System of Logic, Ratiocinative and Inductive : Being a*

Connected View of the Principles of Evidence and the Methods of Scientific Investigation, Toronto ; London : University of Toronto Press.

Nagel, Thomas (1974) "What Is It Like to Be a Bat?", *The Philosophical Review* Vol. 83, No. 4 (October), pp. 435-450.

Nagel, Thomas (1997) *The Last Word*, Oxford ; New York : Oxford University Press.

Nagel, Thomas (1979) *Mortal Questions*, Cambridge ; New York : Cambridge University Press.

Nagel, Thomas (2012) *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*, New York ; Oxford : Oxford University Press.

Northcott, Robert (2013) "Verisimilitude: A Causal Approach", *Synthese*, Vol. 190, No. 9 (June), pp. 1471-1488.

Okasha, Samir (2002) "Darwinian Metaphysics: Species and the Question of Essentialism, *Synthese*, Vol. 131, No. 2 (May), pp. 191-213.

Papineau, David (2008) "Kripke's Proof That We Are All Intuitive Dualists", https://sas-space.sas.ac.uk/892/ . Accessed on 27 October 2019.

Parfit, Derek (1971) "Personal Identity", *Philosophical Review*, Vol. 80, No. 1 (January), pp. 3-27.

Pessin, Andrew and Sanford Goldberg (2015) (eds.) *The Twin Earth Chronicles*, London ; New York, Routledge. First published 1996.

Place, U.T. (1956) "Is Consciousness a Brain Process?", *British Journal of Psychology*, Vol. 47, No. 1 (February), pp. 44-50.

Popper, Karl (1959) *The Logic of Scientific Discovery*, London : Hutchinson.

Popper, Karl (1989) *Conjectures and Refutations: the Growth of Scientific Knowledge*, London : Routledge & Kegan Paul. First published 1963.

Psillos, Stathis (1999) *Scientific Realism: How Science Tracks Truth*, London : Routledge.

Putnam, Hilary (1967) "Psychological Predicates", in W.H. Capitan and D.D. Merrill (eds.) *Art, Mind, and Religion*, pp. 37-48. Pittsburgh : University of Pittsburgh Press.

Putnam, Hilary (1975) *Mind, Language and Reality*: Philosophical Papers, Volume II, Cambridge : Cambridge University Press.

Putnam, Hilary (1979) *Mathematics Matter and Method*, Philosophical Papers, Volume I. London ; New York : Cambridge University Press. First published 1975.

Putnam, Hilary (1981) *Reason, Truth and History*, Cambridge ; New York : Cambridge University Press.

Putnam, Hilary (1982) “Three Kinds of Scientific Realism”, *Philosophical Quarterly*, Vol. 32, No. 128 (July), pp. 195-200.

Putnam, Hilary (1988) *Representation and Reality*, Cambridge, MA ; London : MIT Press.

Putnam (2010) *Meaning and the Moral Sciences*, Oxford ; New York : Routledge. Kindle version. First published in 1978.

Quine, Willard Van Orman (1964) *Word and Object*, Cambridge, MA : M.I.T. Press. Paperback edition. First published 1960.

Quine, Willard Van Orman (1969) *Ontological Relativity and Other Essays*, New York ; London : Columbia University Press.

Robertson, Teresa and Philip Atkins (2018) “Essential vs. Accidental Properties", *Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/essential-accidental/. Accessed on 27 October 2019.

Rosenberg, Alex (2012) *Philosophy of Science: A Contemporary Introduction*, New York : Routledge. Third edition. Kindle edition.

Russell, Bertrand (1905) “On Denoting”, *Mind*, Vol. 14, No. 56 (October), pp. 479-493.

Russell, Bertrand (1910–1911) “Knowledge by Acquaintance and Knowledge by Description” *Proceedings of the Aristotelian Society*, Vol. 11, pp. 108-28.

Russell, Bertrand (1919) *Introduction to Mathematical Philosophy*, London: George Allen and Unwin; New York: The Macmillan Company.

Russell, Bertrand (1927) *The Analysis of Matter*, London : Kegan Paul, Trench, Trubner and Co.

Russell, Bertrand (1948) *Human Knowledge: Its Scope and Limits*, George Allen and Unwin, London.

Salmon, Nathan (2003) “Naming, Necessity and Beyond”, *Mind*, 112, No. 447 (July), pp. 475-492.

Salmon, Nathan (1982) *Reference and Essence*, Oxford : Basil Blackwell.

Schwartz, Stephen (2002) “Kinds, General Terms, and Rigidity: A Reply to LaPorte”, *Philosophical Studies*, Vol. 109, No. 3 (June), pp. 265 - 277.

Searle, John (1958) “Proper Names”, *Mind,* Vol. 67, No. 266 (April), pp. 166-173.

Searle, John (1980) “Minds, Brains and Programs”, *Behavioral and Brain Sciences*, Vol. 3, No. 3, pp. 417-457.

Searle, John (1983) *Intentionality: An Essay into the Philosophy of Mind*, Cambridge :

New York : Cambridge University Press.

Sankey, Howard (1997) "Incommensurability **:** The Current State of Play", *Theoria*, Vol. 12, No. 3 (September), pp. 425-445.

Shaffer, Jerome (1961) "Could Mental States be Brain Processes?", *Journal of Philosophy*, Vol. 58, No. 26 (21 December), pp. 813-822.

Shapere, Dudley (1998) "Incommensurability", in Edward Craig (ed.) *Routledge Encyclopedia of Philosophy*, vol. 4, pp. 732-736. London ; New York : Routledge.

Sharrock, Wes and Rupert Read (2002) *Kuhn: Philosopher of Scientific Revolution*, Oxford ; Malden, MA : Blackwell.

Shoemaker, Sydney (2011) "Kripke and Cartesianism" in Alan Berger (ed.) (2011), pp. 327-342.

Smart, J.J.C. (1959) "Sensations and Brain Processes", *Philosophical Review*, Vol. 68, No. 2 (April), pp. 141-156.

Soames, Scott (2003) *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*, ebook. Oxford Scholarship Online. First published 2002.

Soames, Scott (2011) "Kripke on Epistemic and Metaphysical Possibility: Two Routes to the Necessary A Posteriori", in Alan Berger (ed.) (2011), pp. 78-99.

Stanford, P. Kyle and Philip Kitcher (2000) "Refining the Causal Theory of Reference for Natural Kind Terms", *Philosophical Studies*, Vol. 97, No. 1 (January), pp. 97-127.

Sterelny, Kim (1983) "Natural Kind Terms", in Pessin and Goldberg (eds.) (2015), pp. 98-114.

Stoljar, Daniel (2017) "Physicalism", *Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2017/entries/physicalism/. Accessed on 11 November 2019.

Tobin, Emma (2010) "Crosscutting Natural Kinds and the Hierarchy Thesis", in Beebee and Sabbarton-Leary (eds.) (2010), pp. 179-191.

Walker, Eileen (2012) *The Pathway to Natural Kind Essentialism*, Thesis for the degree of PhD, Department of Philosophy, University of Reading.

Wikfors, Åsa (2010) "Are Natural Kind Terms Special?" in Beebee and Sabbarton-Leary (eds.) (2010), pp. 64-83.

Wilkes, Kathleen (1988) *Real People: Personal Identity without Thought Experiments*, Oxford : Clarendon Press.

Wittgenstein, Ludwig (1968) *Philosophical Investigations*, Oxford : Blackwell. First published

1953.

Wittgenstein, Ludwig (1969) *The Blue and Brown Books*, Oxford : Blackwell. First published 1958.

Woodward, James (2003) *Making Things Happen: A Theory of Causal Explanation*, Oxford ; New York : Oxford University Press.

Yablo, Stephen (1993) "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research*, Vol. 53, No. 1 (March), pp. 1-42.

Yablo, Stephen (2000) "Textbook Kripkeanism & the Open Texture of Concepts", *Pacific Philosophical Quarterly*, Vol. 81, No. 1 (March), pp. 98–122.

Zemach, Eddie (1976) "Putnam's Theory on the Reference of Substance Terms", in Pessin and Goldberg (eds.) (2015), pp. 60-68.