The University Of Sheffield.

A Defence of the Theoretical Relevance of the Term 'Concept'

Bernardo Pino Rojas

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
Department of Philosophy

May 2016

Abstract

The notion of a concept has been widely viewed to be fundamental to understanding the mind. However, some have recently questioned the explanatory role of this notion, asserting that we should eliminate it from our considered theory of the mind. In doing so, these critics are said to endorse a form of concept eliminativism. In this thesis, I challenge concept eliminativism and advance a defence of the theoretical relevance of the term 'concept'.

Firstly, I develop a new general taxonomy of eliminativist arguments and claims through examining a range of different eliminativist projects in different domains. Particularly relevant for this thesis, the proposed taxonomy allows for the characterisation of a type of eliminativism that appeals to the theoretical inadequacy of concepts that do not clearly designate a single class of things.

Secondly, I challenge what is currently the most prominent eliminativist proposal regarding concepts, namely, Machery's concept eliminativism. I begin by providing an overview of contemporary theories of concepts and their main problems. Then I go on to show that Machery's eliminativist proposal fails because it inherits many of the same problems facing the theories of concepts that Machery criticizes. Moreover, I contend that Machery's alternative to concepts is ill-equipped to solve the problem of intentional content. I conclude that these are good reasons to reject the claim that the benefits of eliminating the notion of a concept overweigh the cost of keeping it.

Finally, I defend the theoretical term 'concept' by sketching an approach to natural kinds suitable for an immature science, such as the contemporary science of the mind. I examine several apparently incompatible attitudes towards natural kinds within philosophy of science and argue that this apparent incompatibility demands revision. I address this challenge and develop a positive view that vindicates the scientific relevance of the term 'concept'.

Acknowledgements

I would like to begin by thanking the National Commission for Scientific and Technological Research (CONICYT) for funding my PhD. Without their financial support, my doctoral studies overseas would not have been possible.

Next, I would like to thank my supervisors, Stephen Laurence and George Botterill. I will always be indebted to them for their enlightening guidance, unwavering dedication and endless patience during my research.

Many people have helped me at different stages of a process that started a long time ago in a country far, far away. Some of these people have inspired me profoundly.

I am deeply grateful for the support and education I received from Professors Guido Vallejos and Lorenzo Magnani in Chile and Italy, respectively. Without their support and education, I fear, reaching this point in my academic life would have been an accident.

Special thanks to Bernardo Aguilera, Cristina Roadevin, Emanuele Bardone and Tommaso Bertolotti. In one way or another, they have helped to pave the way for others like me to walk forward. I am very grateful to all my friends here in Sheffield who have made my life in the UK one of my most cherished memories.

Finally, I would like to thank my family for their unconditional support. Above all, I am extremely grateful to Claudia and Constanza for their commitment and tolerance. I am happy to continue with this Voyage in a Parachute.

Table of contents

Word count: 67,430

Preface

The topic of this thesis falls within the realm of the philosophical problems that emerge from contemporary cognitive science. More specifically, it critically examines scientific eliminativism regarding concepts, that is, the claim that the theoretical term ‘concept’ should be eliminated from scientific theorising about the mind. The aim of this work is to challenge what is currently the most prominent eliminativist proposal with respect to concepts and defend the theoretical relevance of the term ‘concept’.

Cognitive science has traditionally been characterised as a constellation of disciplines occupied with answering questions about the nature and the workings of the mind. A common assumption underlying the answers provisionally given to these questions is that the mind is a causal-explanatory factor of intelligent behaviour. The present work is committed to this assumption and the classical understanding of the mind that derives from that interdisciplinary field. According to this understanding, the cognitive mind is best understood in terms of complex representational structures in virtue of which mental processes are computationally implemented.

Indeed, the notion of representation is one of the basic constructs of cognitive science. Theories of concepts arising from this emerging field of empirical study have centred around the view that concepts are the key structural constituents of the complex mental representations involved in our cognitive processes. In this sense, an account of the nature of concepts is deemed a central challenge in practically any theory of cognition committed to the aforementioned classical understanding of the mind.

For over half a century, research in concepts has been marked by debates about the representational nature of concepts and the cognitive functions they may help to explain. Philosophers and cognitive scientists have offered different theories of concepts and there is significant agreement about the explanatory strengths that each of these theories has with respect to a number of cognitive phenomena. However, none of these theories have proved satisfactory for providing an account of all the properties concepts are expected to have and the phenomena they are expected

to explain. Given this scenario of apparent stagnation as regards the development of a unified theoretical framework for the study of concepts, it would appear that, since there is so little agreement about what concepts are, the field is ripe for a radical reform. According to a proposed reform, cognitive scientists should renounce a unified theory of concepts and adopt an alternative that does not in any way appeal to 'concept' as a relevant theoretical term that designates some putative psychological kind. In other words, the proposal is to eliminate the term 'concept' from the theoretical vocabulary of scientific psychology. In the present work, I take issue with this radical proposal and explore the possibility of a different option that vindicates the theoretical relevance of the term 'concept' without risking accusations of scientific stagnation.

There are several motivations for exploring this latter option. First, the notion of concept is considerably useful. Even though there is not yet agreement as to which of the available theories of concepts (if any) should prevail, the notion of concept is directly connected with focal issues in a variety of disciplines such as, for example, linguistics (e.g., the meaning of words), developmental psychology (e.g., conceptual change in human development), cognitive psychology (e.g., categorisation and inference), and philosophy of science (e.g., accounts of natural kinds; cognitive views of scientific change). It is thus reasonable to think that it is good for scientists to keep theoretical terms which are so useful, even if we do not have a very precise understanding of what they might actually designate. Second, it is unclear whether we should apply the same standards for the elimination of theoretical terms to all sciences, without distinguishing between more and less immature scientific disciplines. If the elimination of natural-kind terms is committed to standards for the reality of a given putative natural kind that are too high, then we might find ourselves forced to give up many other terms, irrespective of how useful they are in promoting fruitful enquiry. At the same time, if we are to keep highly imprecise terms merely due to their pragmatic role in promoting further enquiry, then it will be better that we count on a clear idea as to the conditions those highly vague yet useful terms should meet in order to avoid elimination.

A third motivation for attempting to vindicate the term 'concept' involves the productive interaction between theorists working in different areas of cognitive science, including philosophers with an interest in the scientific study of the mind.

Prima facie, it is not very difficult to note that different theorists working in concepts might not be talking about exactly the same thing when they discuss concepts. In particular, this consideration applies to a certain recurrent distinction between psychological and philosophical theories of concepts, where it is claimed that philosophers' hypotheses about concepts are both controversial and irrelevant for psychologists' actual empirical work on concepts. As I will intend to show in this dissertation, this is one of a number of assumptions that motivate a series of unproductive tensions between apparently incompatible attitudes and perspectives towards the study of concepts. It is then interesting to explore a possible re-interpretation of this incompatibility and the conditions for a productive interaction between apparently inconsistent explanatory interests.

My proposal to vindicate the theoretical relevance of the term 'concept' will ultimately consist in a re-interpretation of the interaction between two broadly construed attitudes towards natural kinds: one which is primarily methodologically-motivated and one which is primarily metaphysically-motivated. If we are to come closer to an understanding of the productive role of the term 'concept' within the current study of mind, we need to cope with the fact that the lack of consensus among different theories of concepts is not enough reason for concept elimination. Highly immature sciences must indulge in certain latitude with respect to an account of natural-kind terms. If this is correct, the idea of accommodating highly different attitudes towards the putative natural class of concepts is a perfectly coherent option to explore. The current dissertation is an attempt to deal with this challenge. In what follows, I will first summarise the general topics of each of its six chapters and then I will briefly specify some basic theoretical commitments that this thesis will take for granted.

Chapter 1 explores eliminativism from a broad perspective and provides a comparative picture of a variety of eliminativist projects in different domains. Here the aim to show that eliminativism is a label used for a family of related types of eliminativist arguments and claims. Accordingly, a general taxonomy of these arguments and claims will be provided.

Chapter 2 provides an overview of contemporary theories of concepts which serves as context for the eliminativist view to be discussed in subsequent chapters.

Without an account of these theories, it is difficult to arrive at an understanding of what it is that the concept eliminativist is objecting to. Hence, the main focus of this chapter will be on the problem of the nature of the concepts, that is, the answer that these theories provide to the questions of what concepts are.

Concept eliminativism presupposes that there is some inadequate theoretical term that is ripe for elimination but, as I shall argue in Chapter 3, the prospects of such type of eliminativist projects hang on what it is that the eliminativist takes concepts to be.

In Chapter 4, I go on to show that the alternative to concepts previously presented in the case examined in Chapter 3 is noticeably ill-equipped to solve one of the traditional problems theories of concepts have been expected to address, namely the problem of intentional content. I argue that this is good reason to support the claim that the costs resulting from the elimination of the notion of concept are outweighed by the benefits of keeping it.

Chapters 5 and 6 advance a defence of the theoretical notion of concept by sketching an understanding of natural kinds suitable for the current state of development of the study of cognition. In chapter 5, I argue that the mere realisation that the notion of concept fails to clearly pick out a natural kind according to a given conception of natural kinds does not *ipso facto* entail concept eliminativism. An alternative approach to natural kinds in the domain of the mental is proposed. Chapter 6 is a conclusion chapter, where I take stock and lay out my positive view regarding the scientific relevance of concepts.

The present work will take a number of preliminary assumptions for granted. The first assumption involves the central place of the representational-computational model of the mind in a general theory of cognition. According to this model, the mind is both some kind of computational mechanism and a representational system. The assumption that the mind is a computational mechanism amounts to the view that the functional descriptions of mental processes are equivalent to computational processes. In turn, the assumption that the mind is a representational system amounts to the view that cognitive processes characterised as computations are realised over mental representations which are also the vehicles of intentional content (i.e. they are about events or things in the world). Thus characterized, the kind of intentional psychology

this work is committed to should be understood as a scientific attempt to vindicate our commonsensical explanation of human behaviour characteristically couched in terms of folk notions of intentional states such as beliefs and desires.

A corollary of this first preliminary assumption is that, to the extent that qualitative or subjective states (e.g. those typically attributed to the domain of consciousness) are understood as falling outside the boundaries of the computational-representational model of mind, the present work will not consider those states as relevant for the explanation of the behaviour of agents like us.

The second preliminary assumption is a version of physicalism, understood as the view that mental processes and intentional mental states are part of the physical world. Our mental states have the property of being about something else and that "something else" is the world—or, if you want, aspects of the world, including other mental states. If the processes and intentional states that cognitive science postulates are real (i.e., if they are part of the natural order of things), then those mental processes must be realisable by physical mechanisms and those contentful mental states must ultimately involve the kind of physical entities that are the objects of study of some basic science such as physics. According to a common understanding of the relationship between mental states and those basic physical entities, mental states qua mental causes can be characterised in terms of a type-token distinction—to illustrate, when multifaceted musician Sting utters "message in a bottle" ten times in one of his songs, he has uttered ten tokens of the same phrase type. Thus, while mental states (e.g., my desire to finish this thesis in due time) can be viewed as abstract (i.e., non-physical) types, they can also be viewed as tokens (e.g., a particular occurrence of my desire to finish this thesis in due time) involved in mental causation. Hence, since the tokens of mental-state types are genuine physical objects, the type-token distinction is compatible with a physicalist view of mental states and processes.

The third preliminary assumption is scientific realism. According to this view, scientists are justified in holding a positive epistemic attitude towards the reality of the particulars studied by a given science. Scientists develop theories about these particulars on the assumption that, especially in the case of our best theories, the existence and nature of those putative particulars are independent of those theories. An effect of this claim is the view that the classificatory schemes of a complete

science are to map onto the structure of reality. Put differently, this strong view states that the aims of science are true descriptions of the world (e.g., according to a correspondence theory of truth). Even though it is controversial to claim that a given science is (or can be) a current good approximation to a future correct and completed science, it is possible to distinguish a sort of minimal agreement among realist philosophers of science regarding this positive epistemic attitude. According to this agreement, our best theories are aimed at yielding knowledge about the world whose existence is metaphysically independent of the mind of the knowing subject. In the current dissertation, I endorse this agreement.

Finally, the fourth preliminary assumption underlying the present work is the idea that, whatever else cognitive science amounts to be, it is safe to deem it a highly immature science. What exactly constitutes an immature science is a question philosophers of science have been unable to answer successfully yet. Different proposals have been offered in an attempt to distinguish features and stages of scientific enquiry, especially in terms of some possible set of fundamental commitments (e.g., the idea of *a body of prevailing theories* rooted in the logical positivist tradition; the Kuhnean notion of *paradigm*; Larry Laudan's notion of a *research tradition*, Barbara Von Eckardt's notion of *research framework*, etc.) that a given scientific community might share. For the purposes of the current work, I will only presuppose a rather lax yet relatively uncontroversial characterisation of what may count as highly immature sciences. This characterisation can be outlined in terms of a number of general metascientific claims, such as, for example, the claims that: the theories within highly immature sciences are fruitful yet they tend to resist unification; it is untenable that the outputs of highly immature sciences (e.g., their theories, models, hypotheses, explanations, generalisations, etc.) are specified in the form of the finished products of scientific inquiry; highly immature sciences are in need of a sort of research framework capable of accommodating some minimum set of shared commitments (e.g., regarding the domain of enquiry and some basic empirical questions) held by a given scientific community, etc.

Chapter 1

On Eliminativism

1. Introduction

The topic of this chapter is eliminativism. Essentially, eliminativism is the claim that denies the existence of some type of thing in the world. The first thing you notice when you start researching into this topic is that there is extensive literature about eliminativism regarding *mental states* and our commonsensical understanding of the mind—just try putting the term in any search engine on the Internet. On the face of it, one might be led to believe that the scope of eliminativism is exhausted by the intricate (and sometimes fierce) debates about the mind that have been taking place among philosophers and cognitive scientists. It is obvious that, if you are not familiar with those debates, you will find it difficult to understand what it is that, in the first place, there has been so much fuss about eliminating. This chapter intends to take a step back and explore eliminativism from a broad perspective.

As we will see, people can be eliminativists about different things, which may belong in a variety of domains. For example, eliminativists may reject the existence of supernatural beings (e.g., Santa Claus, deities, trolls, fairies, etc.), biological classifications (e.g., species, races, cells, etc.), and artefacts (e.g., chairs, doorknobs, etc.). However, this does not mean to say that people holding eliminativist claims regarding these different types of things are all eliminativists in the same way. Examining different eliminativist arguments, for instance, can show that there are different ways of denying that there are some X's and, therefore, that there are different ways of being an eliminativist about X's.

One general consideration that motivates the present chapter is that, whereas eliminativist claims, in general, seem to be alike in that they involve rejecting some candidate for eliminativism, there is not just one type of argument for eliminativism that applies for all of them. Indeed, it can be the case that arguments in two different domains are the same, while it can also be the case that there are different arguments within a common domain. Showing that this is the case will be the topic of the next

three sections. The rest of the chapter is aimed at evaluating the possibility of systematizing what seems to amount to a family of related types of eliminativist arguments and claims.

2. Examples of eliminativism

Some examples of eliminativism in different domains can offer a first glance at the diverse character of the eliminativist arguments and claims that this chapter aims to examine. Let us then consider the following cases of supernatural beings, biological classifications and artefacts, respectively.[1]

a) The case of demons:

Belief in magic and the workings of supernatural beings is common to all human cultures and ascribing responsibility to demonic powers for causing certain health states and conditions (e.g., blindness, developmental abilities, or diseases such as epilepsy) is a common historical example of pre-theoretical explanations. Together with the expansion of our reliable knowledge about the workings of natural things in the world around us, demonic explanations have become increasingly unpopular. Most of us would be hardly willing to accept the involvement of magical or demonic causes in people's health and diseases at the expense of natural causes as informed by current scientific enquiry. Thus, by denying the existence of demons to explain the exact causes of diseases, as well as any other natural events, many of us are now eliminativists about demons. Consider, for instance, the following remark made by Ramsey (2013) when commenting on eliminativism within the context of theory change.

> The notion of a demon is just too far removed from anything we now posit to explain behavior that was once explained by demonology. […] We dropped demons from our current ontology, and came to realize that the notion is empty—it refers to nothing real. ("Eliminative Materialism," Section 2.2, para. 3)

[1] For clarity of exposition, I will focus on some paradigmatic version of the eliminativist arguments involved in each of these cases.

Underlying eliminativism about demons is a principle that we may call *causal exclusion*. According to this principle, if instances of a particular type of phenomenon are found to have causes of a certain kind, then there will be no causal room left for certain other supposed causes. Someone may go on to claim that entities whose presumed existence is proved unnecessary for explaining natural phenomena are also ontologically suspect. For instance, given some demonic explanation that posits the existence of several supernatural entities in order to account for each of the currently known types of blindness due to infections (there are at least three), those hypothesised entities are candidates for eliminativism because we now can account for the different causes of visual impairment by natural causes (e.g., certain specific contagious microorganisms) without appealing to supernatural beings in ontological realms beyond the natural world (e.g., demons).

Notice that eliminativism about demons by causal exclusion involves calling into question the existence of a putative type of entity (namely, demons) not only due to their explanatory irrelevance but also due to the way in which they are irrelevant. In other words, the claim is that because demons play no explanatory role, since they are *causally* unnecessary, we are justified in inferring that they do not actually exist.

b) The case of chairs:

We normally think of the world we inhabit as populated by familiar objects such as people, buildings, chairs, stars, etc. and it seems reasonable to believe that these familiar objects, together with their familiar properties (e.g., colour, texture, size, shape, etc.), are features of reality in their own right, in that they have their own existence out there in the external world. However, at least in the case of some of these objects, notably objects like buildings and chairs, our intuition above has been directly challenged.

In theorising about the metaphysics of artefacts, for instance, van Inwagen (1990) argues that the putative objects we call, say, a house, a ship or a chair are not really objects that exist in their own right. More specifically, his thought is that what really exists is not what we call artefacts but only the basic particles they are made of.

So he does not deny that there is some physical stuff there where we claim an artefact is but, instead, he just thinks such stuff is not a thing in its own right. This view is compatible with saying that what we call a house, a ship, a chair, or the like, is only some subatomic particles arranged, say, housewise, shipwise or chairwise, etc. Hence, his conclusion that "There are, therefore, no tables and chairs, and there are no other artifacts" (p. 127).[2] In a similar vein, Elder (2007) describes the case of a carpenter who shapes pieces of wood in order to compose a desk and asks us to consider the question "Is it just that certain pieces of wood or bundles of cellulose fibres have gotten arranged differently towards one another, or has some object different in kind from either the pieces or the bundles been created?" (p. 33).

Both Elder's question about the putative creation of a new object from the carpenter's work and Inwagen's explicit eliminativist conclusion about artefacts exemplify the problems regarding the ontological status of artefacts metaphysicians have been interested in. This problem can be characterized in terms of the following puzzle. On the one hand, contemporary metaphysicians have reasons to believe that artefacts are not part of a serious ontological inventory of the world. This claim does not amount to the assertion that there is nothing in the space where, say, a wooden chair is said to be, but, instead, that there are either just some pieces of wood or, even more strictly, that there is just some set of atoms in the void. On the other hand, there is a strong philosophical argument against the proposition that what we claim to be a given wooden chair and the pieces of wood it is made of are in a relation of identity with one another, in that, when confronted to a relatively standard chair (whatever that might be), one feels equally compelled to view it either as a chair (thus, without noticing what the chair is actually made of) or merely as some bundles of cellulose fibres (thus, without noticing that there is a chair there where the bundles are). Consider that the chair and the pieces of wood differ in several ways from one another. For example, just as in the case of Elder's desk, the pieces of wood in question existed before the chair was made, and they may continue to exist even if, as it might happen, the putative object chair was destroyed. Thus, the question arises as to whether the metaphysician is right when she claims that there are no such objects as chairs.

[2] For a related view, see Unger (1979) and for different views that challenge this conclusion, see, e.g., McGrath (2005), Baker (2007) and Elder (2007).

Let us focus on the argument that artefacts such as chairs do not exist on the grounds that what we call a chair does not amount to an object in its own right. As Elder (2007) suggests, a common way to unpack this argument is in terms of certain worries about composition, where the issue to be decided is under what conditions things can compose an object (e.g., van Inwagen 1990). So, for example, if someone claims that a proper object is the result of the combination between the carpenter's intentional arrangement of some pieces of wood and the uses to which people put that physical arrangement, then the question arises as to whether or not those intentional actions and uses are the kind of things that can be said to really compose a new object. The argument can be summarised as follows:

Premise 1:	There are things that can compose an object and things that cannot
Premise 2:	Real composite objects are made of things that can compose
Premise 3:	Putative composite objects such as chairs are not made up of things that can compose
Conclusion:	Putative composite objects such as chairs are not real composite objects

Consider that the eliminativist conclusion of this argument depends on when composition occurs, given the metaphysical assumption that composition occurs in some cases but not in others. So, eliminativism regarding chairs differs from eliminativism regarding demons in interesting ways. Firstly, rejecting chairs involves a putative case where the eliminativists deny that some physical stuff deserves to be taken as being a proper object, while rejecting demons involves a putative case where the eliminativists simply claim that there are no physical instances of a given type of objects. Secondly, whereas rejecting the existence of demons is primarily the result of an epistemological concern (namely, whether demons exist depends on whether they play a causal role in explanation), rejecting chairs is primarily the result of a metaphysical concern (namely, when it is that a given collection of things that really exist within the space where someone claims a chair is deserves to be taken as an object in its own right). Of course, the argument against chairs may also involve epistemological concerns connected with explanatory relevance, but only after it has been already established that chairs are not made up of things that can compose. Finally, the elimination of chairs, but not the elimination of demons, is constrained in a principled way, namely, in virtue of the metaphysical notion of composition and the

conditions for its occurrence. Note that chair eliminativists of this persuasion can be united by their commitment to the metaphysical presumption that the compositionality principle is the case even if they disagree about its conditions for occurrence. Hence, since the eliminativism regarding chairs can be said to be motivated by the violation of a presumed metaphysical principle, I will call it a case of elimination by *strong metaphysical offence*.

c) The case of moral properties and facts:

Many of us think there are beliefs such as, for example, the belief that the peak of Mount Everest is the furthest summit from the centre of the Earth, which can be true or false depending on how things are in the world. In other words, if it is actually the case that the peak of Mount Everest is the furthest summit from the centre of Earth, then the mentioned belief is true. Otherwise, the belief is false. The metaphysical subtext, in this case, is that there really are certain properties and facts about the world (e.g., planets, mountains, the fact that two objects are at a given relative distance from one another, etc.) that can make certain beliefs (e.g., the belief that Mount Everest is the furthest summit from the centre of the Earth) either true or false. Those of us who are committed to the existence of those types of properties and facts are realist about them. By contrast, those who reject the existence of those types of properties and facts are non-realist about them.

Likewise, there are those who can be said to be realist and non-realist about moral properties and facts, that is, the putative kind of stuff in or about the world that is supposed to make *moral beliefs* true or false. Thus, if someone claims that moral beliefs such as, for example, the belief that all human beings are naturally good or the belief that implementing government surveillance is harmful are true in virtue of how the world is, then there is a presumption that they are committed to the existence of certain moral properties and facts such as, for example, the property of goodness that something may have or the fact that a certain act is morally wrong. It follows from this that those who claim that no moral properties or facts exist will also have to claim that either moral beliefs can be neither true or false in virtue of how the world is or, simply, that those beliefs are always false.

Well, some people claim that moral beliefs are always false because no moral properties or facts are really part of the natural world.[3] For the purposes of the present section, suffice it to go over the main arguments for supporting Mackie's thesis concerning the metaphysical status of morality. Mackie (1977) has advanced what may be called moral eliminativism or the claim that no moral facts or properties exist:[4]

> […] what I have called moral scepticism is a negative doctrine, not a positive one: it says what there isn't, not what there is. It says that there do not exist entities or relations of a certain kind, objective values or requirements, which many people have believed to exist. (p. 17)

This claim is specifically about the metaphysical status of moral properties and the like and it does not entail rejecting common sense moral prescriptions. Indeed, Mackie's moral eliminativism is compatible with accepting the usefulness of certain objectivist moral language, including moral judgments with deontological form (e.g., "Governments ought to be ready to help refugees seeking safety"), so long as the explanation of the apparent objectivity and universalizability of the referents of such language does not appeal to the existence of moral properties and the like as part of the *fabric* of the world. The reason for this compatibility is that, while the universalizability of moral judgments could be validated by the existence of moral properties and facts, the converse does not hold. Hence, the validity (or invalidity) of arguments for or against moral judgments can be said to be independent from the validity (or invalidity) of arguments for or against moral reality. As Mackie (1977) puts it, "The assertion that there are objective values […], which ordinary moral judgements presuppose, is, I hold, not meaningless, but false" (p. 40).

Mackie (1977) presents two main arguments against moral properties and facts, which take the form of two arguments for a species of Moral Error Theory—roughly, moral error theorists accept that there are moral claims but deny that they are

[3] For a defence of an opposing claim, see, e.g., Shafer-Landau (2003) and Scanlon (2014).
[4] In this chapter I use 'moral eliminativism' and 'moral properties and facts' in the same way Mackie (1977) uses the terms 'moral scepticism' and 'moral values', respectively. My choice of terms is merely motivated by consistency of exposition throughout this work. Since nothing hangs on which of this terminology is chosen in this section, you may very well take them to mean the same.

actually true in virtue of how the world is. Mackie calls these arguments the *argument from relativity* and the *argument from queerness*, a version of each of which can be summarised in the following way.

In the case of the argument from relativity, it is first stated that, if there were objective values (i.e., if moral facts and properties were part of the fabric of the world), people would tend to agree about their moral views. However, the argument goes on, disagreement and variation in moral views is abundant across and within different societies, classes and periods. Moreover, it is then stated, abundant disagreement about moral views is better explained by people's adherence to and participation in different ways of life, rather than by the existence of objective values. Therefore, as it is concluded, there are no objective moral values.

The argument from queerness is made up of two components, one is metaphysical and the other is epistemological. Given that the epistemological component is dependent on the metaphysical component, I think it is a good idea to combine both parts into a single argument in order to make that dependence explicit. Thus, the eliminativist argument Mackie (1977) defends can be set out in the following way. Firstly, (Premise 1) morality is committed to very strange or bizarre moral properties which we could only track by some very special perceptual or discerning faculty. Hence, (Premise 2) if there were objective values, the world would have to be such that it contains very strange and bizarre entities or qualities which are "utterly different from anything else in the universe" (p. 38) and we would have to possess some very special perceptual or discerning faculty which is "utterly different from our ordinary ways of knowing everything else" (p. 38). But, (Premise 3) very queer qualities and perceptual faculties are not to be taken seriously. Therefore, there are no objective values.

How does moral eliminativism compare with eliminativism regarding demons and chairs? Well, to begin with, whereas rejecting demons involves rejecting the existence of a putative type of objects, moral eliminativism involves rejecting the existence of both a putative type of objective properties attributed to proper objects as well as putative objective fact-like referents of moral judgments. In turn, while eliminativism regarding chairs is committed to a metaphysical principle, moral eliminativism as examined above is committed to the falsehood of objective moral

facts and properties, which rules out their metaphysical reality. Finally, moral eliminativism, but not demons eliminativism (and less clear in the case of eliminativism regarding chairs), [5] allows for an independent account of the objectification of common-sense belief in the corresponding candidate for eliminativism. That is, moral eliminativism allows for there being some kind of objective common-sense moral claims without there being objective moral facts or properties.

The argument for moral eliminativism is a kind of argument which is primarily motivated by the presumption that putative objects whose metaphysics is too strange or confused do not really exist. For that reason, I will refer to these type arguments as eliminativist arguments by *metaphysical vagueness*.

So far, I have briefly presented three different examples of eliminativism which provides us with a general idea of the kinds of issues that may be at stake when someone intends to argue for the claim that some type of things in the world do not really exist. In particular, examples (a), (b) and (c) show that eliminativist claims arise across a wide range of domains. They also show that there is clearly not just one argument at work for all different types of eliminativist claims. In some cases, for example, eliminativists reject the existence of something because it is redundant for explanatory purposes (e.g., demons) or because it is too far-fetched to be taken seriously (e.g., moral facts and properties). In other cases, they argue that a putative class of objects is a candidate for eliminativism on the grounds that the alleged members of that class do not deserve to be considered existing objects in their own right (e.g., chairs and the like). Let us now take a closer look at some arguments for eliminativism.

3. Types of arguments for eliminativism

In this section, I will focus on eliminativist arguments in different domains. No special attention will be paid to objections that may have been made to these

[5] Someone might think that talk of chairs and moral properties are alike because they are confined to non-scientific domains such as folk talk of artefacts and common sense normative claims, but then again, artefacts are sometimes accepted as proper natural kinds (e.g., in the field of paleoanthropology) and, hence, as objects of scientific enquiry (see, e.g., Machery 2009, section 8.2.1).

arguments. The goal is to simply present and clearly illustrate a list of arguments that have actually been given for eliminativism. A discussion on how these arguments compare and contrast will help us advance a taxonomy of eliminativist arguments, which is to be the focus of the last section.

3.1. Elimination of propositional attitudes

A common intuition is that people's behaviour is somehow determined by their inner mental states and processes. This intuition is reflected in our daily mentalistic discourse such as, for example, when I explain my daughter's decision to postpone her planned picnic in the park due to bad weather by saying that she believes the rain would ruin the picnic and desires to have a good time with her friends. The explanatory and predictive powers of this common sense understanding of the way our minds affect our behaviour have motivated the idea that such understanding really embodies a true theory of our mental life. Roughly, a view like this claims that mental states such as beliefs and desires are real inner states with causal powers and that common sense psychology (also known as folk psychology) presupposes law-like generalizations of the following type: If someone desires that X and believes that the best way to get X is by doing Y, then (all else being equal) she will intend to do Y.

Some cognitive scientists (e.g., Fodor 1987) think science will eventually vindicate common-sense psychology and the existence of mental phenomena as described above. Others think otherwise. Consider the following thesis defended by Churchland (1981):

> […] our commonsense 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, rather than smoothly reduced, by completed neuroscience. (p. 67)

In order to justify this radical claim, known as Eliminative Materialism, Churchland puts forward an argument that takes three main steps. In the first step, it is argued that common-sense psychology constitutes an empirical theory the central

posits of which are so-called propositional attitudes. As their name suggests, propositional attitudes are normally characterized as the combination of attitudes (e.g., believing, desiring, hoping, etc.) and propositions (e.g., "there is bad weather"). According to Churchland, the structure of common-sense psychology resembles that of some physical sciences, the difference being the domain of abstract entities over which they quantify. For instance, while law-like relations in mathematical physics exploit numbers, law-like generalizations in common-sense psychology exploit propositions. Recognizing the theoretical status of common-sense psychology, Churchland claims, allows for a plausible explanation of several issues, including, among others, the explanation and prediction of behaviour. He thus concludes that there are good reasons for theorists to take the theoretical status of common-sense psychology seriously.

If common-sense psychology is really an empirical theory, then it is possible for this theory to be refuted and, hence, for its set of theoretical posits to be displaced. Accordingly, in a second step, Churchland argues that common-sense psychology is deeply mistaken on the grounds that it has a very limited explanatory scope (e.g., it provides no accounts of mental illness or learning processes involving infants and other animals); it is a stagnant theory—"The FP of the Greeks is essentially the FP we use today" (Churchland 1981, p. 74)—and it is just as unreliable as other unscientific theories (e.g., alchemy and cosmology); and it is incoherent with the rest of the sciences in that it is not reducible to any other physical science. In his words, "Any theory that meets this description must be allowed a serious candidate for outright elimination." (p. 76). Thus, the third and final step of Churchland's argument consists of the radical conclusion that, because a theory of propositional attitudes is deeply mistaken, we are justified in inferring that its central posits do not really exist.

The eliminativist conclusion regarding propositional attitudes championed by Churchland can be made explicit in the following argument, adapted from Stich (1983 and 1996):

Premise 1:	Propositional attitudes are the posits of a common-sense psychological theory called "folk psychology"
Premise 2:	Folk psychology is a deeply mistaken theory of the human brain/mind because it epically fails to provide a reductive and coherent explanation of the workings of the human brain/mind
Premise 3:	The posits of deeply mistaken theories do not exist
Conclusion:	Propositional attitudes do not exist

As the second premise of this argument suggests, what is at stake in Churchland's argument against the tenability of common-sense psychology is whether or not there are reasons to think that this folk theory is likely to be vindicated scientifically.[6] For present purposes, suffice it to say that, in order that common-sense psychology can be scientifically vindicated, it should be, firstly, a serious competing theory and, secondly, a scientifically fruitful and coherent theory. Thus, because the eliminativist conclusion of the argument hangs on a total failure to meet those requirements, arguments of this kind can be dubbed eliminativist arguments by *total explanatory failure*.

Compared to the three types of eliminativist arguments I have illustrated in the previous section, elimination by total explanatory failure is closer to elimination by causal exclusion in that both of them invoke some explanatory drawback as a crucial argumentative step. For example, while propositional attitudes are said to fail to explain the causes of behaviour, demons are said to fail to explain the causes of certain diseases. These two types of eliminativist arguments are also similar to one another in that they involve candidates for elimination that are said to be part an explanation of a given phenomenon, though perhaps an explanation no longer regarded as correct by current scientific theory. However, eliminativism regarding propositional attitudes and eliminativism regarding demons differ in an interesting respect, namely, in the way they failed to be scientifically relevant. While rejecting demons is a case where the replacing type of entities (e.g., micro-organisms) is supposed to explain the same type of phenomena (e.g., blindness due to infections) formerly attributed to the replaced type of entities (e.g., demons), rejecting propositional attitudes is a case where such kind of replacement is out of the question, given that common-sense psychology is said to have gone utterly wrong.

[6] Indeed, it can be said that, regarding the prospects of scientific vindication of Folk Psychology, the jury is still out. For various defences of Folk Psychology, see, e.g., Horgan and Woodward (1985), Kitcher (1984), Fodor (1987) and Lahav (1992).

Put differently, it would appear that because in both cases there do seem to be phenomena that are in need of some explanation, viz., why people act as they do, and why people fall sick, one should feel inclined to conclude that there is no significant difference between eliminativism about propositional attitudes and eliminativism about demons. However, when taking a closer look at the way in which demons and propositional attitudes are said to fail to be the central posits of explanatory theories, it is clear that (from the eliminativist's perspective) only propositional attitudes and their causal relation to our behaviour can be accepted as the posits of a certain legitimate scientific contender. The reason for this is that psychological explanation in terms of propositional attitudes is meant to provide an empirical and naturalistic account of the causes of behaviour, and so what the eliminativist thinks is that there must be some more adequate explanation of the causes of behaviour, even though what it is is not yet known to science. By contrast, for the eliminativist regarding demons, even though there is room in scientific theorising for episodes that are taken to be instances of demonic possession, there is no room for instances of demonic possession in scientific theorising about the causes of certain illnesses, since demons are not meant to be legitimate objects of scientific study.[7] In this sense, while there could be said to be no reason to posit psychological explanations that appeal to propositional attitudes (e.g., because there might be better empirical alternative), there would seem to be every reason not to posit explanations in terms of demonic possession.

3.2. Elimination of races

Both scholars and ordinary people have customarily appealed to a putative biologically-grounded notion of races to distinguish alleged natural divisions and subdivisions among human beings (e.g., Caucasian, African, Asian, Spaniard, Amerindian, etc.). Thus, one recurrent assumption behind this practice is that these divisions are biologically real, in that they would have biological foundations which are made manifest in physical and behavioural features such as skin colour, eye shape,

[7] Of course, someone might claim that there can be realists regarding both propositional attitudes and demons, but, then again, the distinction between legitimate and illegitimate objects of study remains since only the realist about propositional attitudes is committed to the kind of realism that is scientifically relevant (i.e., one which is aimed at reflecting the casual structure of the world without appealing to supernatural beliefs).

health status, intelligence, etc. Despite this persistent tendency to categorize human beings into different races on those foundations, eliminativism is a widely held view about biological races nowadays.[8]

The question that biological race eliminativists answer negatively is whether there are races, given their informed judgment that there are no biological grounds for distinguishing human beings into distinct and discrete categories of the sort. These grounds have been normally understood in terms of some type of essences, such as, for example, certain genetic properties, given the belief that these essences would determine certain visible physical traits. However, as the current scientific consensus in this respect suggests, there is no evidence to support the view that there is a significant correlation between people who share phenotypic features and any particular biological conception of essences. Indeed, genetic change does not always result in physically visible characteristics and (contrary to popular belief) even the most visible physical traits fail to work as a criterion to distinguish putative racial groups. For example, Zack (2002) has objected to the idea that the skin colour can be used to distinguish racial groups on the grounds that, because people's skin tones vary gradually rather than discretely, it is not possible to say that people with certain skin tone (e.g., white) always differ from people with a different skin tone (e.g., black) in the same way. Moreover, people who are classified into different races can sometimes be judged to differ from one another less than some people who are supposed to be of the same race. Similarly, given the consideration that biological species are sometimes distinguished by reproductive isolation, it might seem plausible that the concept of race should apply to relatively isolated breeding groups such as, for instance, the Amish in America or Irish Protestants. However, as it has been noted elsewhere (e.g., Mallon 2006), the concept of race is typically used to distinguish groups of people that do not conform to the criterion of breeding isolation such as, for example, those characterised as Caucasian, African, Asian, Spaniard, Amerindian, etc. Thus, as the argument goes, because there is no notion of essences that can be said to biologically ground divisions of human beings in terms of visible physical traits, then there are no grounds for the existence of the scientific category of race. If scientific realism is embraced, such as in the case of Zack (2002), then being an eliminativist about the

[8] For arguments defending the existence of races on different grounds, see, e.g., Gooding-Williams (1998), Andreasen (2004) and Sesardic (2010).

scientific category of race amounts to a metaphysical claim regarding the putative category of race, just as with displaced pre-scientific notions such as phlogiston, humors, etc. Indeed, as Zack puts it:

> The case for the scientific nonexistence of biological race is straightforward and consistent with (accepted) scientific cases for the nonexistence of many other things. (p. 7)

At this point, we may already begin to consider how the case of eliminativism regarding races compares and contrasts with the other cases of eliminativism that we have already discussed. For example, when discussing the elimination of demons by causal exclusion, it was taken as a matter of fact that there is nothing in the world that counts as an instance of the type of things that demons are supposed to be. By contrast, rejecting races does not deny that the entities (viz., people) that are labelled "white person" or "black person" exist. Instead, the claim is that our commonsensical way of thinking about races does not correspond to a real category on the grounds that it does not amount to a scientifically grounded category. In other words, while the one claim is that nothing is actually an instance of a putative given kind (viz., the category of demons), the other claim is something like, given certain superficial differences among people, whatever they might be an instance of, they are not an instance of a given putative kind (viz., the category of races). Thus, resorting to the standard conception of a type/token distinction, to say that there are no demons by causal exclusion means that it is irrelevant to assume that there are *tokens* of some type of things. Conversely, to say that there are no scientifically grounded races means that it is unwarranted to assume that there is a certain *type of things* in itself.

Note that this latter claim also applies when contrasting two different arguments. For instance, it would appear that the argument against races resonates with the argument against chairs in that both of them involve the conclusion that something we think is real is really something that should not be taken to be real in its own right. However, this is not the case, since in the argument against races, not in the argument against chairs, the conclusion accepts the type/token characterisation above. More specifically, what a chair eliminativist denies is that there are certain objects

(viz., chairs) that are instances (or tokens) of a certain type of things which is completely different from the type of things that the subatomic particles that make the objects are instances of.

Is the argument against races an eliminativist argument by total explanatory failure just as the argument against propositional attitudes is? In some important respects, the answer to this question is yes. Consider that, if races are to be rejected from scientific taxonomy, it is, presumably, because this putative category fails to be relevant for supporting many scientific inductions. As Zack (2002) points out, the traditional assumption that 'race' is a biological taxonomy typically means:

> […] a set of physical categories that can be used consistently and informatively to describe, explain, and make predictions about groups of human beings and individual members of those groups." (p. 1)

So, the elimination of races could also be conceived of as a case where the existence of a given type of things is denied on the basis of explanatory failure. However, as in the case of moral eliminativism regarding the need to develop some theory of common sense morality without appealing to the existence of objective moral values, race eliminativists do not need to reject what we may call *talk of races*. Recall that in Zack's argument, just as in the case of Mackie's argument, the eliminativist conclusion is negative in the sense that it is about what there is not, not what there is. So, there is still room for what both authors present as some alternative understanding of race and moral talk, respectively. Put differently, what each of them proposes is an account about certain types of things that we know are false but about which it is somehow good to keep talking (e.g. race talk might still be useful to discuss, characterise and resist certain forms of discrimination and attitudes towards specific social groups and communities which we could not otherwise effectively identify).

The notion of *talk of x* I have just introduced deserves further clarification. Eliminativists about a given type of things typically go on to make a further claim which, even though they also label as 'eliminativism', is not a metaphysical claim at

all. To make this point palpable, let us first consider the distinction Mallon (2006) makes between *normative* eliminativism and *metaphysical* eliminativism. As he points out, what eliminativists normally argue from their conclusion that there are no races is that the concept of race should be avoided:

> Typically there is a close association between metaphysical positions on race and normative positions on 'race' talk. Racial sceptics typically hold that the nonexistence of race supports 'race' talk *eliminativism*. Since race does not exist, it would be false and misleading to continue to use 'race' talk as if it does. (p. 526; emphasis in original)[9]

However, as we will see, it is perfectly possible that metaphysical eliminativism regarding X does not support normative eliminativism regarding our talk of X. For example, Mackie (1977) thinks that, because there are no moral properties or facts (or objective moral values in his terminology), morality is not to be understood as the result of discovery, but as something that is to be made. Roughly, the idea is that issues regarding, say, what moral views to adhere to are to be the result of a decision-making process that yields certain principles of conduct for guiding or controlling people's choices of action. In this sense, even though Mackie denies there are objective moral properties, he is, at the same time, committed to the use of moral talk (i.e., the use of moral terms or concepts) that can play a role in evaluating human conduct. Likewise, Mallon's distinction between metaphysical positions on race and normative positions on 'race' talk in the quote above allows us to better understand cases where eliminativists regarding the existence of races can be, at the same time, advocates of race talk. Indeed, the race eliminativism put forward by Zack (2002) exemplifies such a case in that, even though she claims that races of any type do not exist, she also defends a different view of races (viz., a form of racial constructivism) according to which racial categories are socio-culturally constructed. In such a view, whereas the term 'race' is scientifically otiose, it can very well play a meaningful role in promoting, for example, race-based affirmative action aimed at favouring the well-being of members of certain disadvantaged groups.

[9] Mallon uses the term 'racial scepticism' for the metaphysical view that races do not exist at all.

The distinction between metaphysical and normative eliminativist positions I have just introduced will be useful at the moment of drawing some general conclusions about eliminativist arguments, so I will come back to it in the following sections of this chapter. For now, suffice it to conclude our characterisation of the argument against the existence of biologically-grounded races by noting what I take to be one of its most salient features, namely its weak defence of the elimination of racial terms from scientific taxonomy. As I have said before, this argument is about what there is not, in that it denies that there are races in the world, arguing for the scientific elimination of racial terms without defending a scientific alternative. The argument also states, as a central reason for its eliminativist conclusion, that there are no plausible biological grounds for racial groupings, grounds which have been characterised in certain specific ways. However, it is still possible for someone to argue, for example, that there can be natural racial groupings which do not appeal to those particular foundations, but, instead, to other foundations that may or may not be biological. Likewise, it is possible to argue that, just as it is useful for us to retain racial talk for social or political reasons, scientists may very well retain racial taxonomy even if they doubt that there are biologically-grounded racial groupings, so long as race talk can help discover generalizations about, say, human behaviour or the like. Thus, since arguments of this type are essentially meant to support a negative conclusion, I will dub them eliminativist arguments by *weak metaphysical offence*.

3.3. The argument against the innate

Within the context of philosophical psychology, the questions of what innateness is and what it means to say that a given trait is innate are typically said to have no clear answers. Cowie (2009)[10] characterises at least 16 different ways in which the concept of innateness has been understood, which, in her opinion, reflect the state of disarray the folk notion of innateness is in. Since there is no common way in which all these different understandings of the notion of innateness can be analysed, it is not surprising that some people have felt tempted to question its explanatory usefulness and advocate its ostracism from scientific theorising (e.g.,

[10] Unless otherwise indicated, all reference to Cowie's work in this section corresponds to Cowie (2009).

Griffiths, 2002). Cowie offers an instructive discussion on whether there is a plausible case for the eliminativist option with respect to the innate. She dubs this option as the case for *ElimiNativism*. Importantly for the purposes of this section, Cowie's discussion provides us with a general taxonomy of different types of eliminativist views regarding the innate.[11]

According to Cowie, arguments supporting elimiNativism can be initially distinguished in terms of two different kinds of eliminativist projects which she calls ontological and linguistic (or conceptual),[12] respectively. The question the ontological elimiNativist attempts to answer is whether a given trait in the world is innate. The question the linguistic elimiNativist attempts to answer is whether or not the term 'innate' plays a useful explanatory role in philosophical or scientific theorizing about cognition.

While being distinct from one another, these two projects are also related. Thus, just as there can be reasons to think that theorists should stop using a given theoretical term on the grounds that the putative type of objects that the term purports to refer to does not really exist, there can be reasons to continue with its use. Cowie mentions the case of 'centres of gravity' in different contexts and that of 'electron' as referring to pure particles in chemistry, but we have also discussed the similar case of 'race' and its normative relevance. Alternatively, theorists sometimes conclude that some putative type of objects does not really exist from the realisation that the terms used to refer to them (e.g., 'ether', 'soul', etc.) are not fit for a serious explanatory theory. However, there are terms which, even though theorists would deem them not fit for philosophical or scientific theorising, refer to certain types of things whose existence would not be easy to deny. Cowie mentions terms such as 'sock', 'Hummer', 'dirty joke' and 'herb', but we have also discussed the similar case of 'chair' and other artefacts.

Together with the previous distinction, Cowie also distinguishes between three kinds of eliminativist arguments in philosophy, arguments she invokes to evaluate

[11] Even though Cowie (2009) is specifically interested in assessing Stich's position regarding elimiNativism, I will mainly focus on her proposed general taxonomy of eliminativist arguments with respect to the innate.

[12] With respect to elimiNativism, Cowie (2009) makes no distinctions of usage between 'terms' and 'concepts'. Since nothing in this section is meant to hang on this distinction, I will use these terms interchangeably too.

their possible application regarding the innate. She calls them 'Aren't Any' eliminativism, 'Doesn't Work' eliminativism and 'It's a Mess' eliminativism, respectively. Let us have a look at each of them in turn.

Aren't Any eliminativist arguments are said to be aimed at the ontological eliminativist project and their eliminativist conclusions normally derive from the realisation that nothing in the world satisfies the analysis of a given concept. Cowie thinks Stich (1983) provides a suitable example of this first type of arguments, since Stich's argument for the view that the folk psychological concept of 'belief' will not be part of a mature science of cognition derives from the realisation that nothing really satisfies the analysis of such concept. Cowie's reconstruction of Stich's argument against 'belief' can be set out in three main premises and an eliminativist conclusion. The first premise states that a scientific taxonomy of mental states does not admit states that are individuated in terms of content. The second premise states that scientific taxonomy of mental states is inconsistent with the folk psychological concept of belief because this concept is essentially individuated in terms of its content. The third premise states that folk psychological concepts that are inconsistent with a scientific taxonomy of mental states must be eliminated. Accordingly, the conclusion of the argument is that the folk psychological concept of belief must be eliminated.

One way in which this argument could work in the case of innateness, as Cowie notes, is by establishing that there is something that can be said to be essential to the concept 'innate'. Thus, if this was the case, an elimiNativist could attempt to defend the conclusion of Aren't Any eliminativist arguments of the following form:

Premise 1:	A scientific taxonomy of psychological traits T does not admit concepts that are individuated in terms K
Premise 2:	A scientific taxonomy of psychological traits T is inconsistent with the folk concept of innateness because this concept is essentially individuated in terms of K
Premise 3:	Folk concepts that are inconsistent with T must be eliminated
Conclusion:	The folk concept of innateness must be eliminated

Doesn't Work eliminativist arguments are said to be primarily aimed at the linguistic eliminativist project and their eliminativist conclusions derive from the

realisation that the concept under analysis is part of what Cowie calls a bankrupt theory (i.e., a totally inadequate or scientifically useless explanatory theory). This form of argument is exemplified in the argument against propositional attitudes of folk psychology defended by Churchland (1981). I discussed this argument at the beginning of this section as an example of Total Explanatory Failure eliminativism, so I will not reproduce it here. Instead, I will present the form such type of argument could take if it was to apply for in the case of elimiNativism.

First, the elimiNativist would need to argue that the theory of innateness on which the nativist's research program is based can be taken to be an empirical theory. Then, she would have to provide evidence that such a theory is part of a degenerating research program which simply cannot be scientifically vindicated. The final step would have to argue from such scientific failure to the conclusion that the folk concept of innateness should be abandoned. The general argument can be set out as follows:

Premise 1:	The concept of 'innate' is part of an empirical research program
Premise 2:	The research program the concept of innate is part of is a totally degenerate and inadequate scientific program
Premise 3:	The concepts of totally degenerate and inadequate scientific programs should be eliminated
Conclusion:	The concept 'innate' should be eliminated

Finally, *It's a Mess eliminativist arguments* are said to be primarily aimed at the linguistic eliminativist project, but they can sometimes be taken to support the ontological eliminativist project too. Their eliminativist conclusions derive from the realisation that a given concept simply has no determinate analysis on the grounds that it is too vague or confused for useful explanatory purposes. Cowie illustrates this type of argument along the lines of proposals developed by Griffiths (e.g., 2002), Bateson (e.g., 1991) and Mameli and Bateson (2005). The general form of an It's a Mess elimiNativist argument can be basically set out as follows:

Premise 1:	'Innate' is a (hopelessly) mudded and vague term that resists analysis
Premise 2:	(Hopelessly) mudded and vague terms that resist analysis should be eliminated from philosophical and scientific theorising
Conclusion:	The term 'innate' should be eliminated from philosophical and scientific theorising

Just as in the case of Total Explanatory Failure eliminativism regarding propositional attitudes, It's a Mess linguistic elimiNativism can flirt with It's a Mess ontological elimiNativism, but, again, the linguistic eliminativist conclusion of the one does not directly entail the ontological conclusion of the other. Thus, the challenge for the It's a Mess linguistic elimiNativist is to provide some additional premises to the previous argument such that it can be possible to derive the claim that nothing in the world is innate from the claim that 'innate' is unsuitable for scientific purposes. Cowie points out an interesting issue arising from this latter alternative, given that sometimes concepts that are deemed intractably vague and too messy to be taken seriously can, at the same time, play a productive role in a science of the mind. Hence, in her view, this third type of arguments for ElimiNativism, not the others, is the most interesting option. I endorse Cowie's view with respect to the scientific relevance of many vague and imprecise theoretical terms [13] and, for my own classificatory purposes, I will characterise eliminativist arguments appealing to the theoretical vagueness of certain scientific concepts (e.g., memory, gene, centre of gravity, etc.) as eliminativist arguments by *explanatory vagueness*.

4. Types of eliminativist claims

I am now in a position to present a classification of the several types of eliminativist claims I have illustrated and discussed throughout the present chapter. I will do so in a way that benefits from the distinction Cowie (2009) advances between ontological and linguistic eliminativist projects. Roughly, the first project asks whether some type of objects exists or whether a given property is instantiated in the world and the second is aimed at deciding whether a given theoretical term should be used or abandoned for scientific theorizing. Since eliminativists are not always explicit about the precise scope of their claims regarding the previous distinction, it

[13] The motivations for this endorsement are central to the present thesis and will be elaborated later on in Chapters 5 and 6.

will be useful to set out our classification in such a way that it allows us to distinguish cases in which the eliminativist claims are noncommittal to either the ontological or linguistic projects. Accordingly, a general classification of eliminativist claims can be made in terms of the following three categories:

A. Claims which are committed to both ontological and linguistic elimination
B. Claims which are strictly committed to either ontological or linguistic elimination
C. Claims which are permissively committed to either ontological or linguistic elimination

Type A claims are explicitly committed to the elimination of some type of things in the world and the abandonment of the terms used to refer to those things. The claim against propositional attitudes presented in the previous section exemplifies the kind of claims that belong in this category.[14] Another example discussed above is It's a Mess elimiNativist claims regarding the innate when they argue for linguistic elimination of 'innate' and go on to argue for ontological eliminativism about innate properties and the like. Type B claims, in turn, are explicitly committed to only one of the two eliminativist projects and, at the same time, they are neutral or non-committal about their corresponding counterparts. This is the case of the eliminativist claims connected with the other types of eliminativism about the innate (i.e., Aren't Any, Doesn't Work eliminativisms and It's a Mess linguistic elimiNativism) which are primarily aimed at only one of the two projects (namely, ontological and linguistic elimination of X), as characterised by Cowie (2009). The eliminativist claim about artefacts is another example of type B claims, in the sense that it strictly focuses on the metaphysical question. Finally, type C claims can be exemplified by the eliminativist claims about race and moral properties discussed in section 3, both of which are committed to ontological eliminativism yet non-neutral about the elimination of the corresponding theoretical terms ordinarily used to talk about races and moral properties, respectively. In this case, ontological eliminativist claims are compatible with the permissive usage of theoretical terms for normative or unifying purposes. Eliminativism of this type can be found in Cowie (2009), where the fruitful

[14] Another example can be found in Griffiths (1997), regarding our folk concept of 'emotion'.

usage of muddled theoretical terms is illustrated with respect of the historical development of the concept 'gene'.

With this general classification of claims in hand, I will now turn to supply a classification of the different types of eliminativist arguments discussed in this chapter. This new classification will build on Cowie's proposed taxonomy of different kinds of arguments for elimiNativism and the classification of eliminativist claims proposed in this section.

5. A taxonomy of eliminativist arguments

An important concern within contemporary philosophical and scientific theorising has to do with the evaluation of the extent to which the theoretical posits of a given explanatory theory are both suitable for the discovery of new inductive generalisations and coherent with whatever reliable knowledge we may already have. Thus, it seems reasonable to think that the main motivation for eliminativists to argue for the elimination of X is the thought that X fails to be useful for gaining better understanding of the way things really work in nature. In this section, I will use 'theoretical adequacy' as a pivot concept to draw a general classification of eliminativist arguments. For plainness of exposition, theoretical adequacy should be taken as an umbrella concept that stands for the satisfaction of conditions for explanatory adequacy such as the facilitation of inductive generalisations, explanatory unification and the like. The aim is to show how different eliminativist arguments can be taxonomised on grounds of the way they advocate certain theoretically inadequacy, irrespectively of how *good* or *bad* arguments they can be.

Thus, given the assumption that a certain candidate for eliminativism is thought to fail to meet some kind of conditions for theoretical adequacy, different types of eliminativist arguments can be initially sorted out according to three main categories. Call them Eliminativism due to causal/explanatory inadequacy (Causal/Explanatory Eliminativism), heterogeneity (Heterogeneity Eliminativism), and offence to metaphysical presumption (Metaphysical Eliminativism), respectively.

Roughly, *Causal/Explanatory Eliminativism* is intended to jointly generalise from what Cowie (2009) characterises as linguistic Doesn't Work elimiNativism plus

the addition of its ontological counterpart, i.e. ontological Doesn't Work eliminativism. The reason for the inclusion of this latter subcategory is to provide room for distinguishing cases where the eliminativist conclusion of a Doesn't Work argument is said to be entailed by certain causal/explanatory inadequacy. In turn, *Heterogeneity Eliminativism* is intended to jointly generalise from what Cowie characterises as linguistic and ontological It's a Mess elimiNativisms, where the eliminativist that is commitment to one of them is not committed to the other. *Metaphysical Eliminativism* is intended to generalise from Cowie's ontological Aren't Any elimiNativism. Each of the three broad categories proposed here (i.e., Causal/Explanatory Eliminativism, Heterogeneity Eliminativism and Metaphysical Eliminativism) comprises a number of subcategories, where the several types of eliminativist arguments already explored can be allocated. From this point forward, the following summary table might be helpful.

Types of eliminativist arguments			**Types of eliminativist claims**		
Causal/Explanatory Eliminativism	**Heterogeneity Eliminativism**	**Metaphysical Eliminativism**	A Ontological AND Linguistic	B (*Strictly*) Ontological OR Linguistic	C (*Permissively*) Ontological OR Linguistic
Causal exclusion				•	
Total explanatory failure			•		
	Explanatory vagueness		•	•	•
	Metaphysical vagueness				•
		Strong metaphysical offence		•	
		Weak metaphysical offence			•

Table 1: Taxonomy of eliminativist arguments and claims

The category of arguments which are theoretically inadequate as per Metaphysical Eliminativism comprises two subcategories, namely eliminativist arguments due to strong metaphysical offence and eliminativist arguments due to weak metaphysical offence. Strong metaphysical offence eliminativism (e.g., artefact eliminativism) argues that the existence of some type of thing should be rejected due

to the violation of a certain metaphysical principle. These arguments are aimed at ontological eliminativism (e.g., rejecting artefacts because they have no proper parts can be deemed as violating a metaphysical principle, viz., the principle of 'compositionality') but they are neutral about linguistic or conceptual eliminativism (e.g., artefact eliminativists do not explicitly claim that talking about chairs should be avoided in, say, paleoanthropological research). Regarding our previous classification of claim types, eliminativism by strong metaphysical offence should be associated with type B claims. Likewise, eliminativist arguments due to weak metaphysical offence are aimed at ontological eliminativism (e.g., there aren't essential properties for races), but, since these arguments are not explicitly committed to eliminativism regarding the theoretical terms used to talk about the objects that are candidates for ontological eliminativism, these arguments are best associated to type C claims.

The types of arguments which are theoretically inadequate as per Heterogeneity Eliminativism include both eliminativism due to explanatory vagueness and eliminativism due to metaphysical vagueness. Eliminativism by Explanatory Vagueness (e.g., innateness) argues that, because X has no determinate analysis, it should be rejected. Someone endorsing this type of eliminativism may or may not go on to argue that a given vague concept entails the ontological rejection of the type of thing this concept may putatively designate, depending on whether some intermediate premises are provided to show that this might be the case. Eliminativism due to metaphysical vagueness argues that we should reject the existence of a given type of properties, entities or events on the grounds that their existence would require that we previously accept the existence of some certain types of things that are too strange, confused or improbable, given what we know and commonly think about everything else.

Arguments which are theoretically inadequate as per Causal/Explanatory Eliminativism include what I have called eliminativism by causal exclusion and eliminativism due to total explanatory failure. Eliminativism by causal exclusion (e.g., eliminativism regarding demons) argues that X should be rejected because X is ontologically redundant. More specifically, the claim is that X should be rejected because it does not count as part of a scientifically relevant explanatory theory (e.g., one that is committed to scientific realism or one that can be falsified). Eliminativism by Total Explanatory Failure (e.g., eliminativism regarding propositional attitudes)

argues that X should be rejected because X is part of a deeply mistaken theory, yet a scientifically relevant one.

Recall that Doesn't Work elimiNativism, as characterised by Cowie (2009), was primarily aimed at the linguistic eliminativist project (see subsection 3.3 above). I agree with Cowie that, in the case of the innate, her linguistic-oriented characterisation of Doesn't Work arguments is plausible, but I do not think a main focus on this linguistic orientation reflects the scope of similar eliminativist arguments in other areas. A case in point is the very example she provides to illustrate how Doesn't Work eliminativist arguments work, namely the eliminativism regarding propositional attitudes advocated by Churchland (1981). Indeed, in alignment with our previous characterisation of eliminativism by Total Explanatory Failure (see subsection 3.1 above), Churchland does not only argue for the elimination of 'propositional attitude' but also for the non-existence of propositional attitudes.

Hence, regarding our classification of claim types proposed above, eliminativism by total explanatory failure should be associated to type A claims, while eliminativism by causal exclusion is better associated to type B claims, since causal exclusion eliminativist arguments (not total explanatory failure arguments) are aimed at ontological elimination and are not explicitly committed to linguistic eliminativism—perhaps, in the absence of better and more precise terms, there was a time when talk of demons, phlogiston and the like was useful even though their putative referents were considered highly ontologically suspect.

Explanatory Vagueness eliminativist arguments, by contrast, can be associated to any of the three types of claims, depending on whether the evident vagueness of a given theoretical terms is taken to support the ontological rejection of its putative referent. Take the case of memory, for example. As Cowie (2009) illustrates it, the many different mental states and processes that can be associated with the term 'memory' have led memory experts in different fields to be eliminativists about 'memory' yet not about memory. So, in this case, 'memory' eliminativism by Explanatory Vagueness is best associated with type B eliminativist claims. However, when some appropriate premises are included such that the eliminativist argument about 'memory' is also intended to support the rejection of memory, then memory eliminativism by Explanatory Vagueness is best associated with type A. In turn, even

though terms such as ‘memory’ happen to be theoretically vague, some may still deem them useful terms for explanatory purposes (e.g., Hampton 2010; Strohminger and Moore 2010), irrespective of whether their referents are thought not to really exist (or carve nature at its joints). In this latter case, the use of ‘memory’ can be regimented for convenience in certain classificatory schemes, thus meeting conditions for theoretical adequacy such as explanatory unification. This third type of eliminativism by Explanatory Vagueness is then best associated with eliminativist claims of type C.

6. Conclusion

In this chapter, I have provided a provisional general taxonomy for systematising what amounts to a family of related types of eliminativist arguments and claims. The proposed classification is grounded on the analysis of different types of eliminativisms. Surveying eliminativist arguments and claims in different domains has helped us to build a picture of the different ways people can be eliminativist about some type of thing. Whereas all eliminativist arguments are aimed at rejecting some type of thing, they differ from one another in ways that are relative to the kind of things they reject as well as the scope of their related eliminativist claims. In this respect, I have chosen to set out a taxonomical format which benefits from related proposals and discussions developed by Cowie (2009). The result is a general classification that reflects the way Cowie’s proposal could generalise as a taxonomy of most of the available kinds of eliminativist projects.

Ultimately, all eliminativist arguments can be said to support a conclusion relative to the way eliminativists argue that certain candidates for eliminativism fail to be theoretically adequate. In view of that, the proposed taxonomy acknowledges three broad categories of eliminativist arguments. All the different arguments for eliminativism discussed in this chapter were sorted out according to these three main categories. The category of eliminativist arguments due to causal/explanatory inadequacy is meant to generalise from Cowie’s category of linguistic Doesn’t Work elimiNativism, plus the addition an ontological counterpart. The category of eliminativist arguments due to heterogeneity is meant to generalise from Cowie’s category of It’s a Mess elimiNativism, where the linguistic and the ontological

distinction with respect to a given candidate for eliminativism can be interpreted either as strict or permissive. In turn, the category of eliminativist arguments due to the violation of some metaphysical presumption is meant to generalise from Cowie's category of ontological Aren't Any elimiNativism.

Finally, distinguishing between linguistic and ontological subcategories which focus on some kind of indeterminacy of analysis in the category of eliminativist arguments due to heterogeneity (viz., elimination by metaphysical vagueness and theoretical vagueness, respectively) has helped to draw our attention to the role of vague concepts in eliminativist arguments and the conditions for theoretical adequacy. The issue of vague and messy terms (or concepts) that may turn out to be scientifically productive will be an important topic in later chapters of the present thesis.

Chapter 2

Definitional Concepts and Prominent Revisionist Contenders

1. Introduction

Concepts play a central role in our mental lives. It is the received view that they are the constituents of thoughts and they connect our thoughts to the world. Hence, one common agreement among theorists working on concepts is that a good theory of concepts is essential for several important philosophical issues, a claim that some (e.g., Peacocke 2009) have deemed as probably the only uncontroversial philosophical proposition about concepts that philosophers seem to agree. The current chapter is committed to the view that a good theory of cognition depends on a good theory of concepts.

It is a common intuition among philosophers and cognitive scientists that individuals' intelligent behaviour can be explicated by appealing to cognitive processes where concepts play a central role. This intuition has been well depicted by expressions such as "concepts are the timber of our mental lives" (Prinz 2002) and "concepts are the glue that holds our mental world together" (Murphy 2002). However, attempts to specify the role of concepts within a theory of mind have been the subject matter of interesting debates in the fields of Cognitive Science and philosophy of mind. These two fields of inquiry intersect with each other as many theoretical and experimental developments in current cognitive sciences are highly significant to several philosophical problems (Grush 2002). A case in point is the formulation of well informed theories of concepts that clearly explain the task of concepts in *holding our mental world together*.

In the history of philosophy, one traditional understanding of concepts is based on the idea that concepts are definitions. On this view, for example, the concept woman can be identified with ADULT FEMALE HUMAN BEING.[15] The theory that advocates this view is known as the *Classical Theory of Concepts*. Despite its long-

[15] In this thesis, I use capital letters refer to concepts, unless otherwise specified.

standing historical dominance, there are difficulties with this view and revisionist theories have been offered. Thus, in the last half a century, a variety of competing theories of concepts such as the *Prototype Theory*, the *Exemplar Theory* and the *Theory-Theory*,[16] have been developed as alternatives to the Classical Theory.

In this chapter, I will first address the central aspects of the Classical Theory of concepts and provide a succinct characterization of the main views of concepts developed as a reaction to the Classical Theory. Then, I will summarize a number of criticisms these theories are faced with. The conclusion will be that these theories have common problems which are inherent to the very formulation of the notions of concept they advocate. A clear characterization of these problems will be relevant for chapters 3 and 4, where a recent attempt to reform the current study of concepts will be critically assessed.

2. Traditional and revisionist views of concepts

2.1. The Classical Theory of concepts

According to the Classical Theory of Concepts,[17] most concepts are definitions. What this means is that most concepts have a complex representational structure that encodes conditions for their *application*, i.e., conditions for them to mean what they do.[18] More precisely, the definitional structure of a given concept is assumed to be made up of simpler representations, and these representations are said to provide conditions that are individually necessary and jointly sufficient for something to be in the *extension* of that concept. In this case, to say that *X is in the extension of a concept* is the same as saying that *X is something that the definition of the concept applies to*. Thus, for example, if it is definitional that thistles are plants, then it is not possible for anything but plants to be thistles. In turn, if it was the case

[16] For a panoramic view and a general evaluation of main theories of concepts, see Laurence and Margolis 1999; Prinz 2002, chapters 3-4; Machery 2009, chapter 4. For a detailed account of findings in experimental cognitive psychology, see Murphy 2002.

[17] This view is also known as the *Traditional Theory* or the *Definition View*, and it does not strictly correspond to a single, unitary theory, but to a diverse family of theories committed to the idea that concepts have *definitional structure* (Laurence and Margolis 1999).

[18] Concepts apply to things in the world, such that, for example, if I think about Matilda, the newly born dog in my house, I thereby apply the concept DOG to her. Now, suppose my friend takes Matilda to be a cat when he sees her for the first time. In that case, he is applying the concept CAT to Matilda. It follows from here that concepts can be applied correctly or incorrectly.

that thistles were sufficiently defined in terms of IS A FLOWERING PLANT, IS PRICKLY, IS BIENNIAL, and IS COMPOSITE, then those four representations would be encoded in the definitional structure of THISTLE and they would specify all it jointly takes for something to be a thistle.

Given this characterization of the structure of definitional concepts, the Classical Theory is said to provide powerful explanatory resources for a unified account of a number of important phenomena such as concept acquisition, categorization and reference determination, among others (Laurence and Margolis 1999). Thus, a common model of the acquisition of a definitional concept consists of assembling its constituent parts. In other words, acquiring a concept is the same as constructing a complex representation from its constituent representations, on the understanding that these constituent representations provide conditions that are individually necessary and jointly sufficient for the application of the complex representation. For example, if the concept RIVER has a definitional structure, then learning it consists of learning the definitions of the constituent concepts (e.g., IS A NATURAL STREAM OF WATER, FLOWS IN A DEFINITE CHANNEL, etc.) that provide the conditions for something (e.g., the Po) to be in its extension.

Similarly, the Classical Theory of concepts provides a compelling model of categorization. In general, categorization is understood as the capacity to identify the category to which a given object belongs. Accordingly, a typical model associated with the theory in question consists of checking whether the features *contained* in the concept are satisfied by the object to be categorized. In this case, the term "features" corresponds to the representations that compose the concepts. Thus, for example, someone might categorize something (e.g., the Po) as a RIVER by noting that it satisfies the features that compose the concept river (e.g., IS A NATURAL STREAM OF WATER, FLOWS IN A DEFINITE CHANNEL, etc.).

Analytic inferences are also explained plausibly by appealing to the definition construct. An inference is said to be analytic when the conclusion is guaranteed by the premises supporting it. Consider this example: "My wife is a female parent, so she is a parent." In this case, the conclusion "she is a parent" is true in virtue of the premise "My wife is a female parent." Similarly, a sentence or statement has been characterized as analytic in cases when its truth is necessitated by the meanings of

their constituent elements (Laurence and Margolis 1999). In the example given, the constituent elements are terms such as ‘female’ and ‘parent’, and their meanings are presumed to be contained in the concepts that they express. All this amounts to the thesis that, in those cases, the connections that take place in inferential transitions are intrinsic to the concepts that enter into those inferences (Fodor 1998). Hence, definitions qua mental representations that encode necessary features provide a plausible construct to infer truths from constituent concepts.

One important phenomenon that a good theory of concept is expected to account for is reference determination. The concept RIVER does not mean what one chooses it to mean. A theory of concepts should therefore explain how it is that concepts can have referential properties. According to the Classical Theory, a concept refers to those things that satisfy its definition (Laurence and Margolis 1999). In other words, a concept refers to those things that satisfy the conditions encoded in its complex representational structure. It follows from here that only those things that satisfy these conditions can be in the extension of a concept. In other words, it is the structure of a definitional concept that norm the fact that, for example, only thistles and rivers can be in the extension of THISTLE and RIVER, respectively.

The referential properties of a concept are said to be among its most essential properties. In the case of the Classical Theory, the account of reference determination mentioned above also plays a unifying role, given that the explanation of other phenomena such as categorization and concept acquisition are based on the very structure of concepts that determines its reference. However, in spite of its explanatory power, this theory also faces a number of important problems, and there is now vast evidence regarding certain cognitive effects that this traditional view is not able to predict (e.g., Murphy 2002).

In view of that, a number of revisionist approaches have been developed. Some of them fall under the label of Prototype Theory, which makes reference to the emergence of theoretical and empirical developments in psychology since the 1970s (e.g., Posner and Keele 1968; Rosch and Mervis 1975; Rosch 1978; Hampton 1979). A second group of revisionist approaches have centred around the idea that concepts can be identified with sets of exemplars (Machery 2009), where an exemplar can be roughly understood as a kind of mental representation that is based on memories of

individual category members that we form when we encounter them (e.g., Medin and Shaffer 1978). Finally, there is a third group of approaches that reject the idea that concepts are merely based on superficial similarity to representations of either idealized or actual individual category members. Proponents (e.g., Murphy and Medin 1985) of this alternative group of approaches claim that concepts are theories, a construct that is normally associated with the way scientific knowledge is characterized. This latter set of approaches is known by the name of Theory Theory of concepts. What follows is a brief characterization of each of these three revisionist theories.

2.2. Revisionist theories of concepts

2.2.1. The Prototype Theory of Concepts

According to the Prototype Theory, concepts have statistical structure that encodes the properties (or features) deemed to be possessed by the objects in their extension. The most popular way of characterizing Prototypes as mental representations is in terms of a set of typical feature representations (Prinz 2002). In general, what these feature representations represent are statistically frequent features that members of a category are taken to have.

Put differently, the structure of a prototype concept is said to represent the features that subjects in experimental settings will tend to judge as among the most frequent features in a given category. To illustrate, suppose some of the most frequent features in the category of birds comprise [+feathers], [+beak], and [+ability to fly]. In this case, for example, robins can be taken to be more prototypical of birds than ostriches, given that ROBIN, as compared to OSTRICH, share more feature representations with BIRD. Notably, the feature [+ability to fly] that tends to be part of the concept ROBIN will not tend to be represented in the statistical structure of OSTRICH.

Unlike definitional concepts, prototypes do not comprise necessary and sufficient conditions for the application of a concept. Rather than that, prototypes encode the representation of properties of objects that can be graded statistically. This is why robins can be judged to be a kind of bird more readily than ostriches, precisely,

on the basis of the degree to which these category members tend to possess properties specified by the prototypical representation for birds.[19] On the Prototype Theory, the application of a concept depends on the satisfaction of a sufficient number of features, some of which are considered to have more significant weight than others (Laurence and Margolis 1999). In this sense, it can be said that grasping a concept with statistical structure is *knowing* what features typical things falling under its extension normally have (Fodor 2003). An important assumption among theorists advocating this view is that the mind can group objects to reflect natural groupings in the world, on the basis of similarity relationships that hold both in the world and in the mind (Couchman et al. 2010). However, as it will be discussed later on in section 4, it should not follow from here that relevant empirical research is explicitly committed to this kind of metaphysical assumptions at all times.

Prototypes can also be understood as "summary representations" (Smith and Medin 1981), in that they specify a set of properties or features the sum of which is normally exhibited by category members collectively, but not necessarily individually. What this means is that prototypes can sometimes specify features that are not necessarily co-instantiated in actual category instances (Prinz 2002)—that is, features that are not necessary conditions for category membership. To illustrate, imagine the case of a kind of tree that bears edible fruit that may either be blue in autumn or else pink in spring, with no significant statistical variation in the prevalence of these features (i.e., the features in question have "the same right" to be specified by the corresponding prototype, statistically speaking). Since people may abstract a prototype for this tree that includes both feature representations, it is reasonable to predict that such idealized prototype would allow experimental subjects identify a possible rare species with both features more readily than any other nonprototypical instances. As Prinz (2002) notes, what this implies is that ideal cases could outperform less typical cases, even if these less typical cases, as opposed to ideal cases, have actually been experienced.

One important antecedent to the Prototype Theory are the experiments carried out by psychologists working in different research areas during the 1960s (e.g.,

[19] Some prototype theories characterize prototypes in terms of the representation of properties that objects either possess or do not possess, such as *having wings*. Others characterize prototypes in terms of the representation of properties that objects possess to some degree, such as *being sweet* (Machery 2009, section 4.2).

Posner and Keele, 1968) and early 1970s (e.g., Heider 1972). Results from this work suggested that people store information about categories in the form of representations of what is taken to be their best instances. These contributions preceded the development of the notion of prototype in terms of a "set of typical features" (e.g., Rosch and Mervis 1975). The notion of prototypes can, thus, be understood in different ways and there is currently no clearly specified understanding that all theorists agree with. Notwithstanding this situation, an understanding of prototypes in terms of a summary representation of typical features is often preferred for the purposes of general exposition (e.g., Murphy 2002). This characterization will be helpful to describe how concepts with prototype structure are said to account for phenomena such as concept acquisition and categorization.

A model of concept acquisition that is relevant for the Prototype Theory consists in a mechanism of assembly of feature representations. In other words, one acquires a concept by means of assembling its feature representations. In that sense, concept acquisition is similar to the model of acquisition of a definitional concept. However, since concepts in the Prototype Theory are made up of feature representations that can be graded statistically, the mechanism of assembly in the case of concepts with prototype structure is only aimed to monitor features that tend to co-occur. By contrast, acquiring a definitional concept requires monitoring features that always co-occur, given that each of those features provides necessary conditions for the application of a definitional concept. The Prototype Theory rejects these necessary conditions and considers the representation of a list of statistically prominent feature representations to be sufficient for concept application. Thus, the acquisition of a concept with prototype structure can also be viewed as an assembly mechanism of statistically prominent feature representations that tend to be sufficient for its application.

Categorization is probably the epistemic phenomenon for which the Prototype Theory offers the most natural account. In general terms, this account has been characterized as a similarity comparison process between two representations, where one is for the target category and the other for an instance (Laurence and Margolis 1999). One central idea underlying this comparison process is that category instances are normally taken to be more typical than others depending on the number of prototypical features they have. Indeed, feature or "attribute listing tasks" are common

procedures used by prototype theorists to determine category membership (e.g., Hampton 1979; Smith and Medin 1981). Suppose experimental subjects are expected to judge the similarity between c1 and c2 in terms of the corresponding sets of features F1 and F2 associated with them in that order. A possible account of a similarity comparison process involving F1 and F2 is that these subjects rate similarity in terms of the weight they attribute to a number of overlapping prototypical features. Hence, a possible account of categorization involving, say, items c1 and c2, is that people make use of similarity comparison processes in terms of, say, F1 and F2, to determine the similarity between these items and the prototype for a given target category. In effect, Rosch & Mervis (1975) show that the probability of an item to be judged by subjects as a member of a given category correlates with the degree of similarity between the item and category prototypes.

2.2.2. The Exemplar Theory of Concepts

Not long after the emergence of the Prototype Theory, a new competing theory of concepts was developed. This theory is known as the Exemplar Theory. In this theory, the central idea is that concepts are sets of exemplars (Machery 2009). The question of what exactly an exemplar is has not been clearly answered by proponents of this view (Murphy 2002, p. 58), but it is clear that this view rejects the idea of a representation that can encompass an entire concept. Rather than that, what constitutes our concept of trees, for example, is the representation of the set of trees that we remember. It is an empirical fact that, sometimes, similarity to categories instances which have been previously *experienced* may be better predictors of speed than similarity to category prototypes.[20] This has led some cognitive scientists to conclude that the information about particular category instances that people store is what people use to make their categorization judgments (Prinz 2002). Authors refer to the mental representation of these instances as "exemplars". Hence, a relevant assumption is that it is the exemplars of the objects we remember (more or less accurately) that

[20] Measuring experimental subjects' "reaction time" during cognitive tasks is a typical experimental procedure used to test theories' predictive power in cognitive psychology. The central idea that motivates the methodological interest in this measurement is that the speed of a person's reaction to a given stimulus can provide insights into the nature of the processes underlying the tasks in question.

we consult when we make judgements about a particular category. Consider the following quote:

> Suppose you see a new animal walking around your yard. How do you decide that it is a dog, according to this view? […] Basically, what you do is (very quickly) consult your memory to see which things it is most similar to. If, roughly speaking, most of the things it is similar to are dogs, then you'll conclude that it is a dog. (Murphy 2002, p. 49)

In this quote, the central aspect of exemplar-based categorization is illustrated. According to the Exemplar Theory, we accomplish categorization by comparing a probable category instance to sets of exemplars we store in our memory. Hence, there are, at least, two relevant factors involved in how we categorize an instance: the number of exemplars that an item reminds us of, and the degree of similarity between a given object and each of the sets of exemplars that we remember. Obviously, what an item reminds us of can be a very open-ended set of exemplars (e.g., events can remind us of certain objects or vice versa). For this reason, just an in the case of the Prototype Theory, models of similarity calculation play a central role in this theory.

In a well-known model of categorization proposed by Medin and Schaffer (1978), it is essential to quantify both the importance of the dimension in which a certain set of items can be compared with one another and the amount of similarity in a given dimension. In other words, the model comprises two main stages. In the first stage, we need to determine, on the one hand, matching and mismatching features that the items under comparison may have, and, on the other, how important for determining similarity the dimension of each of these matching and mismatching features is. For example, slave insurrections and revolutionary wars have things they generally share (say, e.g., suffering population; violent uprising against authority; elimination of dominating groups; etc.) and things they do not generally share (say, e.g., introduction of new political doctrine; complete transformation of society; substitution or modification of existing constitution; etc.). Thus, for determining similarity, certain dimensions (e.g., social and economic deprivation) maybe considered to be more or less important that others (e.g., property damage and administrative polices).

In this model, context is said helps in the task of determining the exemplars that are relevant for categorization. In short, consider two sets of exemplars, 'ExC' and 'ExS', where ExC corresponds to a number of contextually selected exemplars, say, the exemplars of the suricates a tourist has seen in the Kalahari Desert, and ExS corresponds to a number of stored exemplars of other categories under consideration, say, the exemplars of wild animals that person has experienced in the past. According to the model in question, the likelihood that something will fall under a given category is said to be a function of how similar it is taken to be to ExC relative to how similar it is taken to be to ExS. In the example given, the likelihood that the tourist categorizes a given animal as a suricate (e.g. while visiting Australia) is a function of how similar that animal is taken to be to the exemplars of suricates she has seen in South Africa relative to how similar it is taken to be to the exemplars of the wild animals she has experienced in the past.

The second stage of the model, in turn, involves deciding how mismatching the mismatching features are. This procedure is aimed to determine the degree of similarity between mismatching features on a given dimension. To do this, Medin and Schaffer's model proposed a *multiplicative rule*. The proposed rule contrasts the additive rule that prototype theorists used (e.g., Tversky 1977). Whereas the additive rule consists in summing shared features, the multiplicative rule consists in multiplying numerical values corresponding to the degree to which each feature is shared (Prinz 2002). The use of this new rule, which has been described in detail elsewhere (e.g., Murphy 2002; Machery 2009; Prinz 2002), provides a bias in favour of cases where an object is taken to be very similar to few exemplars over cases where it is taken to be similar to many exemplars. This is particularly interesting in that the model proves adequate to account for cases in which one can easily categorize highly atypical category members, i.e., category members which are not prototypical.

Thus, for example, imagine there is a political leader of a given ideological persuasion (e.g., a communist leader that successfully advocates a peaceful and democratic transition to socialism) that is very similar to one you have seen before – that is, to your stored exemplar of this leader—, but hardly similar to any other political leaders of the same persuasion (e.g., communist leaders successfully advocating proletarian revolutions). In this case, the use of a multiplicative rule would give the atypical item that is highly similar to your stored exemplar a higher rating

than it would get if one were to determine the similarity between the same item and the prototype of the corresponding category by means of an additive rule. The reason for this is that the model for similarity calculation that prototype theorists use focuses on the similarities that an item shares with most of the members of the corresponding category. Thus, with an additive rule, prototype theories fail to predict cases where atypical items outperform prototypical items.

Another aspect in which the Exemplar Theory is said to outperform its predecessors is the account it provides for concept acquisition. Neither the acquisition of prototypes nor the acquisition of exemplars consists in acquiring necessary defining characteristics of concepts, and both prototypes and exemplars can be learnt by neural networks. However, the acquisition of exemplars, as opposed to prototypes, does not involve monitoring feature representations that tend to co-occur in order to compute the statistically prominent representations that determine the central tendency of a category. As Prinz (2002) notes, evidence also seems to support the prediction that exemplar theorists make about concepts that represent categories which are not linearly separable. For a category to be linearly separable, it should be possible to partition it so that members and non-members are clearly separated.

Sometimes, this partition can be ambiguous or fuzzy. Consider, for example, the possible outcome of cognitive tasks where experimental subjects are confronted with the categorization of items such as vegetable-like fruits (e.g., tomatoes), fish-like mammals (e.g., whales), or atypical birds (e.g., penguins). Exemplar theorists predict that there should be no difficulty in learning categories that are not linearly separable. This prediction contrasts that of prototype theorists. Making use of their models of categorization based on similarity calculation between items and prototypes, prototype theorists have explained why people categorize certain category members (e.g., robins) as birds more rapidly than other category members (e.g., penguins). Hence, their prediction is that linearly separable categories should be easier to learn. However, Medin and Schwanenflugel (1981) have shown that these categories are not learnt more readily than those which are not linearly separable.

2.2.3. The Theory Theory of Concepts

There is a third prominent group of revisionist approaches to the study of concepts that is collectively known as the Theory Theory of concepts. This theory is built, in part, on the basis of the difficulties that both the prototype and the exemplar theories have in order to account for certain categorization judgments. These categorization judgments seem to resort to a type of knowledge that is characteristic of scientific theorizing (Laurence and Margolis 1999).

Given that scientific theories are generally assumed to explain phenomena, instead of providing mere descriptions of them, theory theorists have focused on the idea that concepts must be constituted by knowledge that can help us explain why something happen (Machery 2009). This knowledge has typically been characterized as causal, nomological, functional and the like within a given domain. When the relation between this knowledge and concepts is assumed to be one of identity, then the central idea of the Theory Theory of concepts is that concepts are mini theories of the categories that they represent (Prinz, 2002). Alternatively, this relation can also be assumed to be one where concepts are merely influenced by the kind of explanatory knowledge in question (e.g., Murphy and Medin, 1985). In this case, the identity of concepts is said to be determined by the role they play within a theory (Laurence and Margolis 1999). This role licences the inferential relations in virtue of which concepts are individuated. What this implies is that concepts cannot be understood in isolation from everything else (Murphy 2002), given that the content of concepts (e.g. beliefs about causal mechanisms and general knowledge about the world) is determined *holistically*, that is, by its role in a specific theory. From here, an important implication is that the content of a concept qua theory is not determined by the content of its constituent concepts, such as, for example, in the case of concepts qua definitions.

Not all theory theorists characterize the relation between theoretical knowledge and concepts in the same way,[21] but there are some recurrent claims that they tend to agree with. These claims focus on a number of theoretical facts and highlight the alleged superiority of the Theory Theory of concepts over its immediate contenders (e.g., prototype and exemplar theories). Prinz (2002) summarizes these

[21] For example, some claim that concepts *are* theories (e.g., Rips 1995) and others claim that concepts are *affected* by theories (e.g., Murphy and Medin 1985). For further discussion on this distinction, see Prinz (2002) and Machery (2009).

claims as follows: 1) that concepts encode information that allows us to account for the relations between the features that, according to similarity-based theories of concepts, concepts represent. Such information may provide, for example, an account of why the features HAS WINGS and FLIES encoded in the prototype for BIRD tend to co-occur (e.g., having wings can help to explain the ability to fly); 2) that this information is unobservable. For example, it has been shown that, despite the superficial transformation of a given animal or substance, subjects continue to identify it as the same animal or substance, respectively (e.g., Keil 1989; Gelman and Wellman 1991). What this suggests is that the concepts we use for, say, categorization purposes may comprise hidden features and beliefs about category members; 3) that concepts divide into several domains. Consider, for instance, the *essence* associated with concepts in different ontological categories. In this respect, the case of natural kinds contrasts with the case of artefacts. An artefact that is significantly modified in its appearance may be categorised differently when the modification alters the function it serves (e.g., an old doorknob repurposed as a coat hanger); and, finally, 4) that conceptual development goes through stages that resemble the stages of theory change (e.g., accumulation of counterevidence, theory adjustments, etc.). In this respect, it has even been argued that scientific theorizing and cognitive development are analogous and share the same cognitive mechanisms (e.g., Gopnik 1996).

The four main claims above amount to a general contention according to which concepts embody background knowledge about the world that similarity-based theories of concepts fail to account for. Consider concept learning (e.g., Carey 1985; Keil 1989; Spalding and Murphy 1999; Pazzani 1991). Learning a concept involves the participation of prior knowledge, in the sense that learning a category is facilitated when people have the appropriate knowledge. Experimental evidence for this includes computational models describing the influence on concept acquisition of prior knowledge and learning strategies stored in our memory (e.g., Matsuka and Sakamoto 2007), and neural network model simulations that show how complex concepts with similar features are learnt easier when they are logically consistent with concepts previously learnt (e.g., Hume and Pazzani 1995).

Similarly, the theory theory may explain categorization in terms of, for example, how our beliefs about explanatory relations can affect the way we categorize (e.g., Medin 1989; Murphy and Medin 1985). For instance, consider the possible case

of a person who wins the National Lottery jackpot twice in a row. The reason why we may categorize her as a lucky person does not have very much to do with similarity comparison mechanisms operating between the person's features and the features of the category *lucky*. From the alternative perspective of the theory theory, there must be some background knowledge about, say, the chances of winning the lottery that somehow explains why we take that person to be lucky.

Another advantage of the theory theory over other theories is that it provides a realistic account of categorization, in that the intuition that the concepts involved in categorization already include hidden beliefs about category members is in line with people's tendency towards essentialist thinking (Laurence and Margolis 1999). Consider the case of a dog with his hair dyed reddish-orange and a pattern of dark vertical stripes painted on it. On the basis of Keil's (1989) findings, one could hypothesise that people would still take this animal to be a dog, when asked to make category decisions. If this was the case, then people's tendency to identify such animal with a dog, instead of, say, a tiger, would be better explained in terms of resorting to a mentally represented theory, rather than in terms of quickly monitoring salient observable features. The assumption is that this theory includes the representation of things as having an essence, and that resorting to this kind of information provides a realistic account of psychological essentialism.[22]

The Theory Theory has been usefully summarized as being based on two main ideas, namely, that concepts are some kind of knowledge (or set of beliefs) that underlies explanation, and that they are organised in domains (Machery 2009). These two claims, in turn, are compatible with two prominent developments of the theory of concepts in question. According to one of them, our concepts qua psychological theories are fundamentally analogous to scientific theories (e.g., Gopnik 1996; Gopnik and Meltzoff 1997). Thus, concepts can be viewed as a set of beliefs about categories, where these beliefs can be of an essentialist or an explanatory character. Since these beliefs are organised in domains, concepts can also be viewed as elements or constituents of theories, on the assumption that concepts qua mini-theories are embedded within larger theories.

[22] *Psychological essentialism* is the term Medin and Ortony (1989) coined to describe people's tendency towards essentialist thinking.

Alternatively, according to a second view, our background explanatory knowledge about the world constrains the kind of causal, functional, and nomological knowledge that influences our concepts in a given domain (Murphy and Medin 1985). For example, suppose our psychological theory of terrestrial animal locomotion specifies that if something has long legs, it runs fast. Hence, when developing our mini-theory of, say, poisonous spiders, the mini-theory of terrestrial animal locomotion will cause us to include certain information about the anatomy of arachnids.

To conclude this section, it is fair to say that, just as in the case of other revisionist proposals, theory theorists have managed to raise interesting challenges for the work of philosophers and cognitive scientists trying to specify the nature and the role of concepts in our mental life. Despite this contribution, there are problems and criticisms that all the theories discussed so far are faced with. This is the topic of the next section.

3. Problems of the traditional and revisionist theories

3.1. Why concepts are probably not definitions

The Classical Theory of Concepts (CT, for short) suffers from serious problems. In view of that, the Prototype, Exemplar and Theory theories (PT, ET and TT, respectively) have been developed to replace it. Contrary to CT, these three revisionist theories benefit from the experimental evidence that their advocates have used in order to support certain central theoretical tenets, as presented in the previous section. However, none of these theories have provided adequate and convincing alternative explanations of the main flaws normally attributed to CT. In this sense, PT, ET and TT have their own serious problems, some of which can be said to be common to all of them. Bearing this consideration in mind, I will now discuss the most prominent problems and criticisms of the theories presented in the previous section. As we will see, most of these problems are related to the assumption that concepts must be some kind of complex mental representations.

One of the most evident problems connected with the notion of definitional concept that CT advocates is that most of the concepts we normally use are

tremendously hard to define. It is practically impossible to find a plausible definition for almost any concept, no matter how intuitive a definition may appear to be. In fact, any attempt to specify the necessary and sufficient conditions of a definitional concept seems to be faced with counterexamples. Consider, for instance, the paradigmatic case of KNOWLEDGE as JUSTIFIED TRUE BELIEF. Well known counterexamples have shown that in many cases, even if a belief is both true and justified, one would hardly allow it as a case of knowledge (e.g. Gettier 1963). Other recurrent challenges to the notion of definitional concepts involve questioning definitions such as BACHELOR as UNMARRIED MAN, as well as discussing the difficulty to define concepts such as GAME.

In the case of BACHELOR, it is unclear or controversial that the proposed definition applies, for example, to the Pope and other catholic priests (e.g. Prinz 2002). Put differently, given that the definition of a concept fixes its extension, the assumption is that the concept BACHELOR (or UNMARRIED MAN, for that matter) and its definition UNMARRIED MAN (or BACHELOR) are coextensive (Fodor et al. 1980). However, whereas the Pope and other catholic priests can be part of the extension of UNMARRIED MAN, it is controversial that they are also part of the extension of BACHELOR.

In the case of GAME, in turn, as it was championed by Wittgenstein (e.g., 1953/1958), the features we normally use when we try to define this concept do not seem to be necessary or sufficient for something to fall in the category of games. Consider Hopscotch, Hide and Seek, Simon Says, Solitaire, Telephone, and Freeze Dance. What are the shared defining features of these games that also apply to all and only games? Admittedly, this line of criticism does not prove that there are not any definitions or defining features for a concept like GAME (Murphy 2002, following Smith and Medin 1981), but it does raise certain issues that may critically challenge the classical view in question. One of them has to do with the building blocks of definitional concepts.

As Armstrong et al. (1983) suggests, experimental subjects tend to specify concepts in terms of features that are less primitive or more complex than the concept they describe (e.g. look up the word “simple” in the dictionary). This empirical fact supports the view that concepts can’t be definitions, on grounds of two conceptual

requirements. The first is that the myriad of concepts populating our mental life should result from the combination of simpler concepts (e.g., Fodor 1998; Prinz 2002). Indeed, the number of thoughts and beliefs we can entertain and understand is boundless. This property is known as the *productivity* of our cognitive capacities. A plausible explanation for this phenomenon is that thought is *compositional*,[23] in that the complex concepts that constitute our theoretically limitless variety of novel thoughts are formed as a function of a finite set of constituent concepts plus a finite set of rules of combination. The second requirement is that the extension of definitional concepts must be ultimately fixed by the interpretation of its basic constituents, i.e., those primitive and undefinable components that connect the lexical elements of definitions to the world. Because CT fails to account for this second requirement, it also fails to provide plausible grounds for the first.

It has been argued that the appeal to the definition construct may be based on an overrated emphasis in the *psychological reality* of definitions (Fodor et al. 1980), a criterion that is normally used for the evaluation of psychological theories of concepts.[24] One reason for this is that psychological theories (especially, regarding mechanisms for concept learning) aim to explain the performance of individuals on cognitive tasks by appealing to mental processes where concepts play a central role. And given that cognitive processes can be characterized, in general, as inferential processes, the inferences that take place in virtue of concepts require that these concepts have certain degree of complexity in order to justify the inferential transitions between concepts (Fodor 1998, chapter 4). However, even if the psychological reality of a given notion of concepts is taken to be an important methodological ideal, a theory of definitional concepts cannot simply bracket an account of what the building block of complex representations are. In other words, it is a serious problem for CT to be unable to explain how the proposed notion of concepts can allow for the participation of primitive constituents which are interpreted independently of the definition construct. A better theory should be able to satisfy

[23] For Fodor (1998), this property of concepts is one of the *non-negotiable conditions* that a good theory of concepts should be able to meet. For Prinz (2002), it is one of the *desiderata* that a theory of concepts should ideally explain. In any case, this is a property of cognitive processes that theories of concepts must confront sooner or later.

[24] In very blunt terms, a psychological theory of concepts is said to violate this criterion when it fails to explain aspects of a given phenomenon that are supposed to fall within its explanatory scope.

both the ideal of psychological reality and an account of the basic constituents of thoughts that ground the way thoughts are connected to the world.

So the problems for CT are theoretical as well as empirical. Theoretical problems are mainly connected with the idea that the complex structure of concepts must be definitional because that is the only option that can explain phenomena such as conceptual analyticity (see 2.1., above) and the inferential connectedness commented in the previous paragraph. Regarding analyticity, the problem for CT arises as a result of W. V. O. Quine's criticism of the central role that analyticity had played in logical positivists' project to account for people's purported a priori knowledge (Laurence and Margolis 1999). Quine (1951/1980) critically examined the assumption that there are some truths that are free from empirical revision, and criticized positivists' attempts to define the notion of analyticity for being unclear and circular. He concluded that the distinction between analytic definitions that are true in virtue of the meaning of its constituent terms and those that are true in virtue of how the world is was false because no belief is independent from experience. In other words, there is no such distinction as analytic/synthetic truths because there are no analytic definitions at all, a claim that threatens CT as a theory that provides a plausible explanation of inferential transitions as based on the concepts (or meanings) that participate in those inferences.

Considering the findings on categorization carried out during the 70s by E. Rosch and others (see 2.2.1, above), Murphy (2002) regards *typicality effect* and *category fuzziness* to be among the most prominent empirical challenges for CT. Those findings showed that experimental subjects readily rate items as being more or less typical members of a given category (e.g., Rosch, 1973). From the results of categorization tasks where this typicality effect was studied, it was not possible to detect the incidence of definitions that subjects may be using to make decisions regarding category membership. In fact, the probability for an item to be deemed as member of a given category (e.g., the category of fruits, birds, etc.) was higher depending on its *family resemblance* or similarity as specified in terms of common features that the individuals of a given culture have previously attributed to already known category members. The observed fact that category membership was a matter of degree contrasts with the procedures CT offers for determining category membership unambiguously (Medin 1989; Laurence and Margolis 1999). Definitional

concepts are mental representations which are in a necessary and sufficient identity relation with it constituent mental representations. As Murphy (2002) puts it, this means that definitions pick out all the category members and none of the non-members.

Strictly speaking, Rosch's work was not aimed to provide a theory of concepts, but it did support the hypothesis that concepts cannot be definitions. It was later on that theories of concepts motivated by this line of research were developed, namely the Prototype and the Exemplar theories (see 2.1). These theories are based on the distinction between concept and category. Accordingly, while a concept is deemed a mental representation, a category is understood as a class of objects in the world that the concept represents (Murphy 2002). Through probabilistic specifications, these theories have intended to make the notion of concepts fit the notion of category without challenging the fuzziness of categories that Rosch had discovered. Alternatively, Murphy and Medin (1985) think that neither of these two constructs is enough to account for the processes of categorization and concept acquisition. As described in the previous section, their proposed Theory Theory appeals to our background theories about the world to account for relations of coherence between concepts and the features that constitute a concept. We shall now turn to see the problems these theories are faced with in their attempt to replace the definition construct.

3.2. Why concepts are probably not prototypes, exemplars or theories

A common assumption underlying these three contenders is that concepts can be identified with complex mental representations. In general, both prototype and exemplar theorists agree that concepts cannot be definitions. In turn, theory theorists claim that concepts cannot be prototypes or exemplars all along. Nonetheless, prototypes, exemplars and theories are competing constructs with a common motivation, namely, Rosch's work on categorization that challenged the plausibility of definitions as the fundamental construct of a theory of concepts. Issues concerning category fuzziness and the psychological reality of concepts are among the central challenges raised by this experimental tradition.

In this section, I conclude that none of these challenges is adequately met by any of the three constructs in question. I ground this conclusion on two main claims. The first is that a plausible theory of concepts cannot be merely grounded on our categorization performance, which is only one of the cognitive phenomena where concepts are involved. The second is that a focus on the problems of category fuzziness and the psychological reality of concepts is not enough to guarantee the stability or publicity[25] required by (at least some) concepts. Notably, the relative stability of concepts proposed by the similarity-based approaches fails to account for concept publicity, given the epistemic account of content-constituting inferences that these approaches presuppose. I shall provide reasons for thinking that each of these claims is probably true and common to all these theories. Before that, I will succinctly summarize the main problems attributed each of these competing theories of concepts.

3.2.1. Problems of the prototype construct

There are recurrent problems regarding prototype-based representational structures. First, many concepts lack prototypes. What are, for example, the prototypical properties of *bilingual second time mothers*, or *the king of Chile*? In fact, one can possess these concepts without even knowing the corresponding prototypes. Second, concepts that are assumed to have prototype structure fail to cover highly atypical instances and incorrectly include non-instances (Laurence and Margolis 1999). For example, a theory of prototype has serious problems to represent diseases. The perceptual properties of a disease are its symptoms, and symptoms of a disease can hardly be allowed to be constitutive of it. Indeed, a disease can be easily misdiagnosed if it is only judged by its superficial effects (e.g., people with normal blood sugar can experience symptoms of diabetes, and recently diagnosed people with diabetes may report to experience none of the relevant symptoms). A third problem involves the apparent empiricist character of prototype theories, given that they mainly represent perceptual properties. This could explain the explanatory impotence of the theory regarding cases where people can have concepts without knowing its

[25] The publicity of concepts is grounded on the intuition that a same concept can be share by different people and also by a given person at different times. It may be clarifying to assume a type/token concept distinction and suppose that different people can share tokens of the same concept type (see, e.g., Fodor 1998; Prinz 2002).

prototype. However, some think concepts with prototype structure can represent statistically prominent features that are not of a perceptual kind (e.g. Machery 2009). For instance, we may consider certain numbers to be more typical of ODD NUMBER than others or, for that matter, EVEN NUMBER.

Another problem for prototypes is that they cannot provide an account of the characteristic compositionality of our cognitive capacities. Consider, for a paradigmatic example, that the prototype for PET FISH (whatever it is)[26] has little to do with the prototypes of its constituents PET and FISH. This clearly suggests that the prototype of a complex concept is not formed as a function of the prototype of its constituent concepts. Compositionality is an important property of concepts because it allows us to account for the productivity of thought. The notion of concept that the classical theory of concepts put forward was a good candidate to account for this property, in that the definitions of, say, PET and FISH can be assumed to be the constituent parts of the definition of PET FISH. It is, therefore, a problem for prototype theories that the complex structure of definitional concepts they intend to revise does a better job at explaining how we can learn and form concepts such as BILINGUAL SECOND TIME MOTHERS.

Likewise, the model of concept acquisition the Prototype Theory has involves problems that are inherent to its formulation. According to this theory, acquiring a concept (e.g. ROAD) involves assembling a certain set of typical features. However, each of these typical constituent features (e.g., CAR, LANE, ROUTE, TRAVEL, etc.) can also be considered as instances of some category in their own right, so to say. What this suggests is that the statistical procedures involved in concept acquisition should also apply for the particular features that typical category instances normally have. If so, learning a new concept would entail circularity. Consider, for instance, that the feature *lane* can be taken as part of the prototypical representation for ROAD, and that, at the same time, the feature *road* can be taken as part of the prototypical representation for LANE. Hence, the question arises as to whether the acquisition of some set of primitive feature representations can be reduced to a sensory level, such that those primitive features can be identified with sensory properties. If this was the

[26] In this case, PET FISH is an example akin to TROPHY HEAD, where the first lexical constituent item acts as an "identification attribute" (e.g., X that is Y, and, therefore, Y that is X), rather than to COWGIRL, where the first lexical constituent item, COW, acts as a "characterization attribute" (e.g., X that does/has/etc. Y).

case, the Prototype Theory would not have to face the problem of presupposing procedures of concept acquisition that involve representations without statistical structure (e.g., sets of exemplars). However, it would face the empiricist problem that most concepts resist analysis in sensory terms (see, e.g., Laurence and Margolis 1999), which does not render the model of concept acquisition proposed by this theory particularly more compelling than that of the Classical Theory of concepts.

Finally, it is not clear whether or not the very notion of prototype—understood as *statistical knowledge about a category*—necessarily presupposes that of "exemplars" (see Murphy 2002, p. 64) at least in some cases. The idea here is that a prototype view is not entirely self-sufficient to account for cases where individuals form the prototype of a category member they see for the first time. Notably, in these cases it makes more sense to talk about an exemplar-based representation, rather than a statistical summary representation of the category in question. Similar considerations have been invoked to assess possible cases where individuals form the prototype of a single category member that they see on regular occasions. These cases raise additional serious concerns over whether prototypes can be identified with concepts (see Rosch and Mervis 1975, p. 575, for early background on the distinction between typicality manifestation and category representation). Maybe, exemplar approaches provide a better revisionist alternative than similarity-based approaches turning around the notion of prototypes. Let us, then, consider the problems of this alternative option.

3.2.2. Problems of the exemplar construct

As described in 2.1.2., the prototype and exemplar views do not involve *defining* characteristics of the mental representations they identify with concepts, which makes them immune to the problems that the definitional view of concepts faces. However, there is a common problem that some authors have characterized as the selection problem (e.g., Machery 2009; Smith and Medin 1981). According to this concern, individual category members have (indefinitely) many properties, so prototype and exemplar theorists need to explain why prototypes and exemplars represent the properties that they do, instead of others. Whereas neither of these

theorists addresses this issue, the implications of the selection problem seem to be more dramatic in the case of exemplars.

Prototype theorists assume that different statistical parameters in long-term memory may be involved in the cognitive processes underlying different cognitive capacities, such as identification and categorization (Machery 2009). Alternatively, since exemplars theorists claim that exemplars are memories formed of the category members that we encounter, it is the same set of exemplars that have to be involved in the cognitive processes underlying different cognitive capacities. This makes the selection problem more pressing for an exemplar theory of concepts, as it amounts to answer the question of what it is that an exemplar actually represent. In this respect, Murphy (2002) points out the case of an animal you can see just for a fraction of a second, for which reason you are not able to pay very much attention to it. Or else, the effects it may have on someone to encounter the same animal on a daily basis because it happens to be her neighbour's pet. Are these encounters relevant to decide what could count as an exemplar of what? Could a toy dog count as an instance for an exemplar of dogs under any given circumstances? Exemplar-based models of identification and categorization assume that similarity calculation involves the participation of all of the instances encoded into memory. If this is so, plausible models should state where one should draw the line that divides what counts as an exemplar from what doesn't.

In general, exemplar theories of concepts that make use of a similarity calculation model based on a multiplicative rule are more effective at identifying atypical instances that prototype theories fail to account for with their models based on an additive rule (see 2.2.2. above). However, this advantage implies high costs in terms of memory storage and processing, given that we are assumed to store memories of all the exemplars we experience and make use of all of them for similarity calculation at the moment of categorizing objects (Prinz 2002). Likewise, it is unclear how such a maximal amount of exemplars can explain the property of conceptual compositionality, ideally without presupposing further increase of our processing load. One could hypothesise, for instance, that exemplars can combine with one another in terms of certain association procedures. Maybe my exemplar of DIARY automatically retrieves my exemplar of PAPER because of my familiarity with paper diaries. Similarly, my daughter's exemplar of DIARY may readily retrieve

her exemplar of DIGITAL because of her familiarity with digital diaries. However, this does not account for our capacity to entertain novel concepts of category instances that we have not experienced before or, simply, that we are not familiar with, such as VEGETABLE ICE CREAM, PET VOLCANO, or, to invoke a more eminent example, COLOURLESS GREEN IDEAS. Hence, an account of conceptual combination based on the association procedure hypothesised above would be of very limited explanatory scope. But, in fact, even if we grant that this limited proposal is plausible enough to account for conceptual combinations such as PAPER DIARY, exemplar theorists would be confronted—as they are—with a more serious question. Is the exemplar of PAPER DIARY a function of the exemplars of PAPER and DIARY? Clearly, the answer is negative.

Finally, these limitations have implications for explaining other properties of concepts in terms of exemplars. Abstract and fictional concepts (e.g., PEGASUS, ROCINANTE, UNICORN, TROJAN HORSE, etc.) are especially difficult to explain for similarity-based theories of concepts, notably for theories that lack an adequate account of the compositionality of concepts. It can be thought that the content of fictional concepts is explicated in terms of the content of exemplar combinations (Prinz 2002), but, as we have just seen, exemplars can't compose. So, exemplars cannot explain how it is that we can possess abstract or fictional concepts because exemplars are the stored mental representation of category members previously experienced. *A fortiori*, exemplars are the sort of thing that people do not share, given that, if people have their own stored memories of category members, then they have their own exemplars. Thus, identifying concepts with sets of exemplars implies abandoning the possibility that concepts be shared by different people or, even, by the same person at different time, considering that our set of exemplars can be modified every time we experience a category instance. In other words, the exemplar construct fails to meet the publicity requirement that concept theorists consider a good theory of concepts should be able to meet.

3.2.3. Problems of the theory construct

The Theory Theory puts forward the idea that concepts belong to larger bodies of knowledge. This knowledge is assumed to be arranged in specific domains with

explanatory principles that help us make sense of different aspects of the world (Prinz 2002). However, knowledge varies between people and stages in their lives, which makes it hard to allow that people can have the same concept. Some (e.g., Fodor and Lepore 1992; Rey 2009) consider this *variability* to be characteristic of most (if not all) epistemic conceptions of concepts.

In a nutshell, an epistemic conception of the notion of concept corresponds to one that is said to be dependent on 'how we *know* what is X,' instead of being dependent on 'what *is* X' irrespective of the fact that we are able to think about X (see, e.g., Rey 1985). According to Fodor (e.g., 1998), philosophers and cognitive scientists' tendency to assume that the investigation of concepts should begin by enquiring about conditions of concept possession, and, then, by inquiring about the identity conditions of concepts is the current trend in the study of concepts. Fodor has extensively argued that, when this strategy is followed, conditions for having concepts are invariably understood in epistemic terms, given that this methodological trend is normally constrained by a pragmatist approach that focuses on what a person is able to do in order to possess a concept. On the understanding that the empirical study of concepts is inevitably committed to certain preliminary assumption about their nature (Millikan 1998), this approach is said to be in agreement with the metaphysical commitment that concepts are some kind of epistemic capacity (or epistemic "know how"), instead of, for example, abstract entities or mental particulars. A case in point is when possessing a concept is equated with our capacity to draw inferences. In this case, Fodor's contention is that most theorists working in concepts have wrongly taken it for granted that a concept can be unequivocally individuated in terms of its inferential relations to other concepts. On his view, one of the main reasons why this mode of individuation is untenable is that it entails a holistic determination of the content of people's thoughts.

In a holistic theory of content, the content of a concept is determined by its role within a system of beliefs about the world, which, in the case of the theory of concepts under discussion, corresponds to a mental theory. What this means is that the individuation of a concept is determined by its inferential relations to all other concepts in a given person's belief system about the world that a given individual may have at a given stage of her cognitive development. Hence, the content of a concept

can't remain stable across changes in its mental theory,[27] which is why the proposed notion of concept fails to provide a principled basis for comparison. In other words, the theory construct fails to offer a way to compare whether two people (or the same person at different times) share the same concept. This poses a serious challenge for cognitive scientists advocating this construct because they need to count on a stable notion of concept over which psychology can generalize.

In fact, precisely because bodies of knowledge (or systems of beliefs) are unstable, the theory construct is unable to satisfy the publicity constraint, a requirement that is necessary for the generalizations about concepts that could help explain behaviour. Consider, for instance, the case of intentional explanations, where attitudes (e.g., beliefs, desires) are required to be composed of the same concepts in order to explain people's actions. Thus, an explanation such as 'If *S* desires *Q*, and *S* believes that not-*Q* unless *P*, then *S* does *P*' can generalize from particular cases where different people (Ss), actions (Ps) and desired outcomes (Qs) are involved, provided that, in all those cases, people share the same mental states (i.e., propositional attitudes) and, therefore, the same concepts that are its basic constituents.

To put it bluntly, the Theory Theory fails to provide an adequate account of intentional content. It fails to determine reference, which means that they fail to refer to categories. Without any specification of the necessary and/or sufficient conditions for something to be in the extension of, say, APPLE, theories do not contain enough represented information to tell what belongs and what does not belong in the category of apples. Suppose, for instance, there is a theory of apples, according to which apples are the kind of things that grow on apple trees and fall off at maturity. Such description is inadequate to say that the theory in question is a theory of all and only apples. One reason is that the very concept APPLE is being presupposed in the theory that is supposed to explain and, therefore, individuate what APPLE is. Another reason is that apples are not the only kind of things that this theory can apply to. Consider pears, apricots, and even red-eyed tree frog tadpoles!

[27] This is what Laurence and Margolis (1999) have dubbed as the Problem of Stability regarding the holistic determination of the content of a concept.

Similarly, theories may contain false beliefs or incorrect information. Suppose I believe it is the spirit of my ancestors that cause frog calls. If so, and if all the components of a theory are necessary for reference determination, then it follows that my concept FROG would be empty. At the same time, it is patent that holding such false belief about frog calls does not preclude me from entertaining the same concept FROG CALL that people normally entertain when they refer to frog calls without having to assume the existence of spirits of any kind. What this shows is that we can possess a concept even if we are mistaken or ignorant about the (essential) properties that the concept is deemed to possess. This is what some authors have called the Problem of Ignorance and Error (see, e.g., Laurence and Margolis 1999; Rey 1983).

Compositionality is another problem for the Theory Theory. Since the theory construct is explanatory impotent to account for intentionality, the issue of how compound concepts can be formed as a function of its constituent concepts poses a serious challenge. Theory theorists might want to explain intentional compositionality in such a way that, for example, theories can be said to be formed out of its constituent theories. However, this is not a promising strategy for at least two reasons. First, it is the implausible that theories can exhaustively contain other theories, even if we accept the idea that the theory construct can be understood in the form of mini-theories, as it has been characterized by Prinz (2002). Second, we lack an account of how (mini-)theories could combine with one another such that the content of the compound theory resulting from that combination can be said to be directly inherited from the content of the constituent (mini-)theories in question. Recall that the content of a theory is likely to be determined holistically. This implies that theories do not contain other theories exhaustively, which is why it makes no sense to talk about compound and constituent theories the way it is required in order to account for the compositionality requirement. Clearly, my theory of TREE FROG TADPOLE does not inherit all the possible inferential relations that my (mini-)theories of TREE and FROG, individually, may hold with any other set of more or less adjacent theories.

As characterized in 2.2.3., theorists have appealed to the theory construct as a useful tool to gain insight into cognitive development. Laurence and Margolis (1999) have focused on Gopnik (1996) and Gopnik and Meltzoff (1997) in order to illustrate a recurrent view according to which theory change between different stages of development mimics theory change in the history of science. Machery (2009)

summarizes three properties of scientific theories that are relevant to this analogy: introduction of theoretical entities related in a systematic set of law; explanatory and predictive purposes; theory changes in response to evidence. However, theory change in science is still a very poorly understood phenomenon, which is why its alleged mechanisms could hardly be said to play a central role in the explanation of other poorly understood phenomena such as cognitive development and concept learning. To illustrate this point, let us consider the current epistemological debate concerned with the role of models in scientific cognition.

Given the agreeable consideration that an important part of scientific investigation is carried out on models (e.g. Nersessian 2008), there has been an increased interest in studying scientists' modelling practices in order to understand how new scientific concepts are formed and what the mechanisms underlying theory change in science might be. Different approaches in this field focus on theorizing about a wide variety of practices that might support the claim that reasoning is model-based and that the explanation of scientific cognition has to focus on some notion of model-like representation (Thagard et al. 1999). Nersessian (1999), for example, has studied a set of practices that include creating analogies, employing visual representation and thought experimenting. Magnani (e.g., 2001 and 2009), in turn, has incorporated an additional dimension that he calls 'eco-cognitive' according to which concrete manipulations of external objects influence the generation of new hypotheses in science. A common consensus among these researchers has to do with a rejection to the idea that a mere propositional formulation of inferential processes can provide an exhaustive account of hypothesis generation.

This latter formulation has been advocated by an earlier trend initiated by Hanson (e.g., 1958, 1960) and Harman (e.g., 1965) on the basis of further development of Peirce's ideas about abductive inferences.[28] The main project, in this case, involves the view that a logic of discovery, alongside a logic of justification, is possible. Within this trend, research has basically aimed to determine necessary constrains to render abduction a reliable inference, given the substantial assumption that inferential processes can be understood as propositionally-based and governed by

[28] For Peirce, abduction was special non-deductive inference that takes place in the process of forming an explanatory hypothesis. However, the current understanding of abduction centres over the idea of inference to the best explanation. See Douven (2011) for further clarification.

explicit rules or constrains. Alternatively, researchers advocating Model-Based Reasoning (MBR) claim that their proposed *style* of reasoning, where implicit constrains are said to abound, can better explain the nature of ampliative reasoning in science. However, MBR theorists tolerate the possibility that certain aspects of scientific cognition are propositionally-based, which goes to suggest that none of the approaches in question, by themselves, seems to be able to provide a comprehensive explanation of new hypothesis generation.

Admittedly, creative aspects of scientific cognition could be conceived of as some kind of sophisticated version of ordinary cognition. But, given the challenge of developing a common theoretical framework that harmonizes the deeply opposing substantial assumptions in question (viz.., that, in some cases, the format of the representations we manipulate when we reason is model-based and that, in some other cases, it is propositionally-based), it is controversial that current theorizing about scientific ampliative reasoning, in particular, can provide reliable insights about ordinary cognition in general.

Finally, as Prinz (2002) highlights, there is empirical evidence that eclipses many of the advantages normally associated with the Theory Theory. Some of this evidence includes experiments showing that this theory can get categorization judgments wrong (e.g., Hampton 1995; Smith and Sloman 1994), and that beliefs about essences do not necessarily coincide with categorization (Malt 1994). In turn, Malt and Johnson (1992) have shown that certain artefact objects that theory theorists assume to have functional essences are often categorised on the basis of superficial similarities. Admittedly, the Theory Theory can accommodate many of these findings, which somehow suggests that they cannot be taken as decisive evidence against the psychological essentialism that, as the theory assumes, may underlie people's categorization judgments. However, people fail to specify what the essential property of a given category might be, even in experiments where it seems evident that concept users take some hidden essence for granted. Likewise, theory theorists do not provide a convincing characterization of such essences. Gelman et al. (1994), for example, describe the idea of psychological essentialism in terms of people's tendency to assume that there is "a true, underlying nature that imparts category identity" (p. 344). This description may be useful to debunk prototype theorists' models of categorization based on readily observed superficial features of objects. In fact, a

common criticism of this theory is that it does not provide a principled way for selecting those features, except for the claim that it is previous knowledge that renders them salient (see, e.g., Murphy 2002). The notion of hidden essential features that the Theory Theory advocates could solve that selection problem, yet this theory would still need to answer what those features are and what it is that makes them essential.

4. Concluding remarks: common problems among revisionist constructs

Prototype, Exemplar and Theory theories have their own particular problems, but the most important problems are common to all of them. Some of these common problems involve, at least, three requirements that, according to the Cognitive Science community, a good theory of concepts should be able to meet: intentionality, publicity and compositionality.[29] The reasons why these requirements cannot be met by the revisionist theories examined in this chapter are inherent to the formulation of these theories. In identifying concepts with prototypes, exemplars and theories, revisionist theories take it for granted that the way to individuate a concept is in virtue of its relations to other concepts. In this particular respect, none of these theories is more promising than the definitional view of concepts they intend to displace.

According to the Classical Theory, the mode of individuation of a definitional concept is in virtue of its relations to other definitional representations that are said to be the constituents of the concept. In turn, the Prototype Theory assumes that a concept is individuated by a set of statistically frequent relations to its constituent features. In the exemplar theory, a concept is individuated by its similarity relations to the representation of its exemplars. And The Theory Theory puts forward the idea that a concept is individuated by its relations to other concepts in adjacent theories, which, in turn, can be part of a whole theory of the world relative to a person's cognitive development.

However, it is only in the case of a definitional concept that the relation between a compound concept and its constituent representations is assumed to be one of necessity. Hence, the problem of the stability regarding the content of a concept becomes more pressing in the case of the revisionist theories. Indeed, all these

[29] For a broader list and a general characterization of each of these requirements, see Prinz 2002, c. 1.

theories are said to predict both that categories will be graded in terms of typicality and that there will be borderline cases showing that their boundaries are fuzzy, given the evidence for typicality effect (Murphy 2002). Let us consider the implications of the instability in question with respect to the requirements mentioned above.

In order to account for reference determination, similarity-based approaches are sometimes said to refer in terms of the same features by which they categorize (Prinz 2002). This is the reason why prototype theorists sometimes claim that a prototype reference is graded. However, it is hard to allow that the FISH prototype, for instance, refers to salmons or trout more than it refers to seahorses. In other words, just because seahorses are less typical fish it doesn't mean that they are fish to a lesser degree. So, maybe, prototypicality is graded, but class membership is clearly not. Similarly, exemplar theorists defend the idea that concepts can be identified with sets of exemplar representations based on the information about previously encountered instances that people store in long term memory. It follows from this that being a concept is dependent on being something that can be represented by memory instances. However, this excludes concepts that may refer to categories whose category members we have never experienced, such as, for example, PROMETHEUS, UNICORN, etc., even if they result from the combination of the representation of category instances that we may have actually experienced (e.g., ATTACK SHIPS ON FIRE OFF THE SHOULDER OF ORION[30]). In the case of the Theory Theory, in turn, it is part of the very formulation of the theory construct that it does not contain enough represented information to refer to categories, precisely, because the content of a concept is supposed to be determined in virtue of its relations to the role other concepts play in adjacent theories. Such theory of intentional content attainment does not provide a criterion to determine which set of those relations can be relevant for the individuation of a concept.

Another implication of the instability of intentional content that the revisionist theories are bound to is the relaxation of the publicity requirement. The thought seems to be that, if strict content stability is not necessary for psychological explanation, then it must be that similarity of mental content is enough to guarantee that different

[30] "I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I watched c-beams glitter in the dark near the Tanhauser Gate. All these moments will be lost in time. Like tears in rain. Time to die." (The Top 10 film moments, 2000).

people (or the same person over time) can share the same concepts, such that they can be subsumed in the corresponding psychological generalizations. But this is clearly implausible, at least in the case of the theories in question, all of which are committed to the substantial preliminary assumption that the individuation of concepts is metaphysically dependent upon our epistemic capacities to access the objects in their extension. This commitment is not explicit but it is made manifest in the methodological tendency to assume that the way to individuate concepts is via the satisfaction of conditions of concept possession expressed in terms of certain epistemic procedures. As characterized in 2.2. and 3.2., these epistemic procedures can be based on the description of stereotypic features of the objects that may fall under the extension of a concept, and/or the description of reliable cognitive mechanisms that allows for the epistemic access to those objects (e.g., our ability to draw inferences). One clear example is a focus on the possible mechanisms underlying typicality effect, i.e., individuals' tendency to judge category members to be more or less similar to one another. Revisionist theories, especially similarity-based approaches, take it for granted that this cognitive effect can somehow account both for learning differences and reaction time differences in categorization judgments (Murphy 2002, p. 64).

However, in proposing that sharing similar content is enough for people to share the same concept, prototype theorists presuppose content identity. Experimental subjects tend to produce the same features when providing the characteristics of a given category, and this is taken to be evidence that the content of similar concepts is reliably shared when these concepts share a sufficient number of the same features. But this is clearly question-begging, given the implied suggestion that what guarantees the similarity of content at the level of concepts is the identity of content at the level of features (Laurence and Margolis 1999).[31] Moreover, empirical evidence shows that typicality judgments about category instances, and the features that are deemed to be typical of them, vary from person to person and moment to moment (e.g. Barsalou, 1987).

Exemplar theorists are faced with the same dilemma. To the extent that exemplars are said to be represented by a list of correlated features, the problem of

[31] See Fodor and Lepore 1992, for extended discussion.

presupposing content identity is the same as described in the previous paragraph. If, on the other hand, exemplars are represented as entire objects (e.g., abstract, summary representations), the problem of stability is probably intractable. Given that exemplars are supposed to be made out of people's own memories, our concepts of a given category are different by definition, i.e. by reasons that are inherent to the formulation of the proposed notion of concept.

Similarly, the Theory Theory is unable to satisfy the publicity requirement because, by definition, so to speak, the content of a concept is permanently revisable. This theory is based on the assumption that the individuation of a concept (and its content) depends on the explanation of our cognitive capacity to draw theoretical inferences among concepts in different theories. In the context of the Theory Theory, this epistemic commitment implies that the content of a concept is relative to an indefinite set of inferential relations between continuously changing bodies of beliefs. When it is put in that way, this mode of concept individuation entails a holistic determination of the content of people's thoughts and their constituent concepts. Hence, the same variability constraining the failure of the theory to account for intentionality (or reference determination, for that matter) constrains its failure to satisfy the publicity requirement. This variability is inherent to the formulation of the theory construct.

Finally, revisionist theories are embarrassingly poor at dealing with the requirement of compositionality. Theories that identify concepts with prototypes, exemplars and theories presuppose that most concepts require certain degree of complexity and that this complexity is central for psychological explanation. This is the reason why these constructs are individuated in terms of the pragmatic value they seem to have with respect to cognitive tasks where concepts are assumed to be involved (e.g., reasoning, categorization, induction, language understanding, etc.). This is a methodological assumption that may lead us to an ontological claim about concepts according to which concepts *are* complex mental representations whose constitutive features are the inferential relations established between concepts.[32]

[32] This applies for cases where complex concepts are assumed to be composed of other complex concepts as their proper parts, as well as cases where complex concepts may stand in a privileged inferential relation to other complex concepts that are not exhaustively contained by any other concept. For a general characterization of these two models of conceptual structure, see Laurence and Margolis 1999.

However, from the pragmatic value in question, by itself, it does not follow that most concepts are, *ipso facto*, complex or structured mental representations. Neither does it follow from that methodological assumption that the only way to satisfy the identity conditions of a concept is by specifying the constitutive features of a complex mental representation. Consequently, unless a concept is individuated in a way that is not dependent on the multiple inferential relations that may be established among conceptual systems, the problem of content stability is inevitable. This problem underlies the particular difficulties regarding compositionality that the revisionist theories are faced with.

Thus, all the revisionist theories critically examined in this chapter fail to satisfy the requirements under consideration. I have endorsed a characterization of this problem in terms of a metaphysical commitment that identifies the notion of concept with some kind of epistemic "know how" that is common to all these theories. This commitment is said to be the invariable metaphysical subtext of theories that attempt to individuate concepts via their possession conditions, such as when, for example, possessing a concept is identified with constructs that are dependent on our capacities to sort things or with constructs that are dependent on our capacities to draw inferences (e.g., Fodor 1998). If this is right, it can be concluded that the main problems underlying the revisionist psychological theories of concepts in question are inherent to the formulation of the notion of a concept they advocate.

Chapter 3

Against Concept Eliminativism

1. Introduction

Chapter 1 examined eliminativism in different domains and concluded that most eliminativist arguments can be said to support a conclusion relative to the way eliminativists argue that what they think should be eliminated fails to be theoretically adequate. The taxonomy proposed in that chapter included a category of eliminativist arguments according to which a given theoretical term is a candidate for eliminativism when it has no determinate analysis and is too vague to play an adequate theoretical role (i.e. eliminativism by explanatory vagueness).

In this chapter, I take issue with a recent eliminativist proposal regarding concepts put forward by Edouard Machery. On his view, cognitive scientists should abandon the notion of concept because there is something deeply mistaken about the way they have scientifically understood it. Because this proposal ultimately defends the claim that keeping this notion makes communication among cognitive scientists cumbersome, concept eliminativism of this type can be said to fall in the category of eliminativist arguments mentioned above. Here I will focus on Machery's general argument against the notion of concept, arguing both that this argument fails and that we should reject the elimination of concepts from the theoretical jargon of cognitive science. Other more specific drawbacks about this type of eliminativist projects will be developed in the two subsequent chapters.

2. Machery's argument for concept eliminativism

As we saw in Chapter 2, prominent competing theories of concepts have posited different constructs with particular structures (e.g., prototypes, exemplars and theories) as plausible notions of concepts to explain a variety of cognitive phenomena and abilities (e.g., concept learning, inferences, categorization, language understanding, etc.). Advocates of each of these theories have attempted to show that

their proposed construct is better at explaining certain phenomena than those proposed by other competing theories (Prinz 2002; Machery 2009, 2010). None of these theories is entirely independent; and cognitive scientists sometimes agree that different constructs may co-exist (Murphy 2002). Some of them even think it is plausible that concepts can have different types of structures as their components such that these components can be associated with different explanatory roles (e.g., Laurence and Margolis 1999). In any case, a shared assumption is that there must be a single theory of concepts that is able to account for all the relevant phenomena, an assumption that Machery (2009) has called "the received view."

An interesting reaction to this view is Machery's version of conceptual heterogeneity, the view that the class of concepts divides into several distinct kinds that Machery takes to be different bodies of knowledge with very little in common. This view is in contrast with the assumption that the class of concepts form a single natural kind with many common properties, which Machery calls "the Natural Kind Assumption." Denying this assumption is the main motivation behind Machery's argument for conceptual heterogeneity (2009, 2010), which is committed to the alternative view that current psychological theories of concepts should be considered as distinct theories, given that what they are really about is how distinct (or heterogeneous) bodies of knowledge can account for similar but independent phenomena. A relevant presumption here is that, in any given science, theoretical terms that fail to pick out a natural kind should be eliminated from the corresponding scientific classificatory scheme.[33] Machery's hypothesis is based on the claim that the empirical plausibility of the different forms of conceptual representations posited by the most prominent competing theories of concepts (i.e., prototypes, exemplars and theories) violates the assumption that concepts are a natural kind.

It is worth noting that Machery's argument for heterogeneity draws an eliminativist conclusion about the class of concepts from the conclusion that 'concept' as a theoretical term is explanatory inadequate. Prima facie, this claim might resonate with other more traditional eliminativist views about the mental (e.g., Churchland 1981) where it is claimed that certain theoretical posits are both enmeshed in a theory that is but a total explanatory failure and doomed to be displaced by future scientific

[33] See Chapter 5 for a more detailed treatment of this claim.

developments. However, even though Machery thinks the notion of concept is ripe for scientific eliminativism, he does not go on to claim that concept theorists are utterly wrong about their particular theoretical proposals. Instead, his view is that talk of concepts should be avoided for the purposes of explaining many of our higher cognitive competences because there is an alternative psychological theory of these competences that need not make use of the notion of concept. Addressing the main aspects of the argument supporting this view is the aim of the current section.

2.1. The Heterogeneity Hypothesis

According to Machery (e.g., 2005, 2009, and 2010), prototypes, exemplars and theories are all distinct *bodies of knowledge* equally (henceforth, BoKs) relevant for scientific inductive generalizations about the mind. He thinks these bodies of knowledge are distinct because they are sustained by different causal mechanisms that are responsible for the manifestation of different cognitive capacities such as induction, categorization, analogy-making, etc. He maintains that the empirical plausibility of this latter claim provides grounds for the elimination of the theoretical notion of concept. His alternative proposal is based on five main tenets which he jointly dubs the *Heterogeneity Hypothesis* (e.g., Machery 2009, p.4). The general structure of the argument for his hypothesis can be reconstructed as follows:

Premise 1:	If a theoretical notion does not pick out a single homogeneous natural kind, we should eliminate it from scientific taxonomy
Premise 2:	The psychological notion of a concept is used to refer to prototypes, exemplars and theories
Premise 3:	Prototypes, exemplars and theories do not form a single homogeneous natural kind altogether
Conclusion:	We should eliminate the notion of a concept from the scientific taxonomy of psychology

Machery's justificatory support for his heterogeneity argument can be associated with three main argumentative stages. In the first stage, he argues that the explanatory advantages of a number of very different theories of concepts suggest that for each object, substance, event, etc., there can be several concepts. Thus, for example, for the category of horses there can be HORSE1, HORSE2, etc. The aim of

the second stage is to argue that these several concepts are really *distinct* bodies of knowledge used in distinct cognitive processes, which would provide grounds for them to count as distinct natural kinds in their own right. And, finally, the third stage appeals to the claim that a set of distinct (or heterogeneous) bodies of knowledge (notably, prototypes, exemplars and theories) should replace the allegedly misconceived notion of concept as a single homogeneous class relevant for scientific generalizations about the mind.

Let us focus on the first stage. On Machery's view, Prototype, Exemplar and Theory theorists are, on the whole, right about the type of *information* they claim is involved in the processes underlying people's higher cognitive capacities. In other words, they are right when they claim that, in categorizing, drawing induction, etc., people make use of the information about categories that prototypes, exemplars and theories are supposed to encode. If this is the case, then, contrary to the Received View, the explanatory advantages of a theory advocating any one of these three constructs should not be viewed as evidence against theories advocating any of the others. Machery's relevant conclusion is that the empirical plausibility of prototypes, exemplars and theories implies that, for each category, an individual must have several concepts (tenet 1). His proposal is to characterize these several theoretical entities as *co-referential* bodies of knowledge.

Suppose we can have different concepts for the category of birds. Let them be bird-prototype, BIRD-exemplar and BIRD-theory. Allegedly, people can sometimes produce conflicting categorization judgments about the category of birds. This is the case when, for example, we judge a penguin to be a bird if asked to consider the way they are born, but judge it not to be a bird from the set of exemplars of the instances of birds that we remember. In this latter case, we may fail to consider certain knowledge about evolutionary and developmental mechanisms that may have been relevant for the first categorization judgment. According to Machery, conflicting judgments like these could not be possible if the causal mechanisms in the categorization processes where BIRD-prototype, BIRD-exemplar and BIRD-theory are involved were the same. Then, as he contends, it must be that his proposed distinct bodies of knowledge are actually sustained by different causal mechanisms underlying higher cognitive processes (i.e., they are not *coordinated*, to use Machery's expression).

In the second stage of his argumentative strategy, Machery (2009) argues that a variety of empirical evidence supports the claim that the use of each of these bodies of knowledge is actually sustained by different causal mechanisms. Notably, processes where prototypes, exemplars and theories are involved can yield different categorization (or inductive, etc.) judgments. For example, suppose an inductive judgment—e.g. the judgement that *if something is a cat, it will probably purr*—is best explained in terms of processes where prototypes or exemplars are involved (e.g., Sloman 1993; Murphy 2002, pp. 180-190). And suppose also that another inductive judgment—e.g., the judgement that, if your cat catches a cold, it will probably be sick for some days—is best explained in terms of processes where theories are involved. Then the thought is that, in those cases, the causal mechanisms sustaining the processes where prototypes or exemplars are involved are independent of the causal mechanisms sustaining the processes where theories are involved. In this sense, a given category can be said to be represented by co-referential bodies of knowledge that have very little in common (tenet 2).

Given this characterization of distinct bodies of knowledge as sustained by different causal mechanisms, it would seem that Machery's conceptions of prototypes, exemplars and theories correspond to distinct natural kinds. Consider, for example, Richard Boyd's theory of natural kinds. Boyd (1991, 1999) has put forward the idea that, for something to count as a scientifically relevant natural kind (e.g., some biological species), it should possess a typical set of properties sustained or brought about by causal mechanisms. Likewise, Machery posits that prototypes, exemplars and theories are distinct natural kinds because the knowledge activated (i.e., retrieved from long-term memory) in processes where they are involved is sustained by independent causal mechanisms.

Machery admits the importance of distinguishing knowledge that is constitutive of any one of these three distinct bodies of knowledge and knowledge that is not. In this sense, he contends that a necessary condition for these theoretical entities to count as natural kinds is that they be understood in terms of "bodies of knowledge that are used by default in cognitive processes underlying higher cognitive competences" (Machery 2009, p. 11). In this quote, a crucial idea is that of some

information that is used by default.[34] I have previously equated the notion of default with the idea of something being used automatically, but I think a brief characterization of this notion would be useful to understand the centrality it plays in Machery's theoretical proposal.

The notion of *default* stems from the study of default inferences in fields such as artificial intelligence and computer science, where it is hypothesized that certain inferences are normally drawn unless additional information is provided (Machery 2009). Likewise, on Machery's view, the idea that a body of knowledge about *x* is used by default when we reason about *x* (or categorize, make analogies, etc.) amounts to conceiving of this body of knowledge as some kind of defeasible knowledge that we take to be relevant and preferentially available when reasoning about *x* (or categorizing, making analogies, etc.). Thus, the proposal that prototypes, exemplars and theories are distinct bodies of knowledge used in distinct cognitive processes (tenets 3 and 4) is based on the presumption that they are bodies of knowledge about *x* which, out of the whole knowledge about *x* that may be stored in our long-term memory, are preferentially retrieved from memory to be used in independent cognitive processes underlying our higher cognitive competences.

The last stage of Machery's case for the Heterogeneity Hypothesis involves arguing from the claim that prototypes, exemplars and theories pick out distinct natural kinds to the elimination of the notion of concept. A common assumption among philosophers and scientists with naturalistic commitments is that scientific taxonomies correspond to natural kinds in that they actually reflect real categories in nature, subject to subsequent refinement. Within Cognitive Science, theorists working in the development of psychological theories of concepts have generally taken it for granted that the notion of concept reflects a natural kind and can support causally-grounded generalizations.[35] Machery disagrees with this latter assumption and claims that prototypes, exemplars and theories are distinct natural kinds relevant for scientific generalizations about the mind in their own right.

[34] In some more recent work (e.g., Machery 2015), it is claimed that the notion of 'default' should be applied, not the *use* of knowledge, but to the *retrieval* of knowledge from long-term memory. I will not consider this revised conception of default in this chapter, since it is not relevant for the current discussion. However, I will pay attention to this distinction in Chapter 4.

[35] See Chapter 5 for a discussion on the notion concept and its putative natural kind status.

As pointed out in the previous stage, Machery contends that these three theoretical entities do not form a single natural kind because they do not have causal mechanisms in common. Hence, he claims that the failure to recognize this fact explains contemporary theorists' failure to develop a successful unified psychological theory of concepts. In other words, there can be no theory of concepts based on the assumption that prototypes, exemplars or theories form a single natural kind. According to Machery, there is enough evidence to conclude that there can be no theory of concepts based on the idea that the notion of concept is a single natural kind. He therefore recommends that, in order to discover scientifically relevant generalizations, cognitive scientists abandon the notion of concept and focus on prototypes, exemplars and theories as distinct bodies of knowledge.

3. Some problems with Machery's distinct kinds of fundamental concepts

In Chapter 2, I went over the main aspects and criticisms of the Prototype, Exemplar and Theory theories of concepts. The conclusion there was that these theories have common problems, and that these problems stem directly from the fact that these theories individuate concepts by relation to other concepts. In view of that, I argued that prototypes, exemplars and theories fail to provide an adequate account of certain basic properties of concepts.

In this section, I will show how Machery's proposal is vulnerable to the same problems that Prototype, Exemplar and Theory theories are faced with. These are problems that his proposal can directly inherit from other theories. Additionally, I will argue that Machery's proposal faces its own particular problems. By making these problems explicit, I intend to prepare the ground to show that the argument bearing upon the Heterogeneity Hypothesis presented at the beginning of the previous section fails.

3.1. Why distinct bodies of knowledge can't displace concepts

I will now focus on some of the problems of Machery's view of distinct concepts. Most of them can be briefly stated as follows, depending on whether or not

they are clearly inherited from the three theories of concepts that Machery's view is intended to combine.[36]

Problem 1: Distinct bodies of knowledge can't account for Intentionality. Concepts represent things other than themselves, notably they represent things in the world. What they represent is known as the intentional content of the concept and a good theory of concepts should explain how concepts attain their contents. For current purposes, this is the same as explaining how concepts refer to things in the world. The shortest way to show that Machery's theoretical proposal fails to explain how the distinct bodies of knowledge attain their intentional content is to establish that it fails to do so because the Prototype, Exemplar and Theory theories fail to do so as well (See Chapter 2, section 3.2). But Machery's view raises its own particular difficulties in this respect because it is not obvious that his several kinds of concepts can be identified with mental representations of some kind.

The idea of distinct bodies of knowledge about a given category presents Machery's view with the following dichotomy. Either it maintains a commitment to the different types of structured mental representations that the Prototype, Exemplar and Theory theories posit, or it remains neutral about the representational nature of his bodies of knowledge. In both cases, his proposal will be faced with particular problems. On the one hand, if his bodies of knowledge are to be identified with complex mental representations bearing structural relations to their parts as per the Prototype, Exemplar and Theory theories, then his proposal could be said to inherit the very problems of these theories to explain how those representations refer to the things that they do. On the other hand, if Machery's proposal of bodies of knowledge remains neutral about their representational nature, then this proposal faces the additional challenge to provide a suitable theory of reference. This theory should help us understand both how the distinct bodies of knowledge about a given category can refer to things in the world and, at the same time, how they can be co-referential.

According to the Heterogeneity Hypothesis, for most categories, we typically have a prototype, a set of exemplars, and a theory about them. Machery claims that

[36] In this section, I will succinctly enumerate what I take to be some of the main problems of Machery's view of bodies of knowledge. Chapter 4 will provide a more detailed discussion on Machery's view with a focus on the problem of content, a problem I think an alternative to concepts should not disregard.

these entities are distinct but co-referential default bodies of knowledge. However, there are reasons to think it is hardly likely that these three entities can have the same referents (e.g., Margolis and Laurence 2010; Jacobson 2010). For example, a body of knowledge about particular instances of hammers (i.e., a set of exemplars) may apply to an atypical emergency hammer for breaking through the windows on public transport, while knowledge about prototypical hammers may not. Likewise, a body of knowledge about prototypical apples may apply to non-instances of the category of apples, such as some items on a decorative plastic fruit bundle platter, while a theory-driven body of knowledge about causal or functional properties of apples may not. These possible cases are useful to show how the three bodies of knowledge can fail to be co-referential.[37]

The previous examples are also useful to draw attention to the distinction between categorizing and referring. The notions of prototype, exemplar and theory have been useful to develop different models for categorization. In Machery's proposed theoretical framework, they are meant to serve similar purposes, but how successful these constructs are to account for categorization is neither necessary nor sufficient for determining the reference of the distinct bodies of knowledge.

In examining linguistic evidence for the Heterogeneity Hypothesis, Machery argues that two distinct bodies of knowledge about, say, tomatoes can be retrieved by default depending on what people take tomatoes to be in different respects such as, for example, vegetables or fruits (see, e.g., Machery 2009, p.72-73). Specifically, he thinks that, when people judge that the sentence "Tomatoes are vegetables" is true, they make use of a TOMATO-prototype (rather than, e.g., a TOMATO-theory) because they take the sentence to claim that tomatoes look like vegetable. This could help us explain why people categorize something as a member of this or that category, but this explanation is not a good reason to believe that there are distinct co-referential concepts (e.g., TOMATO1 and TOMATO2) because there are different ways in which we take something to be.

[37] Here, I take it for granted that, for different bodies of knowledge to be co-referential, they have to share the *same* content. Someone might say that sharing *similar* content is a sufficient condition for different bodies of knowledge to be co-referential, provided that such similarity is accounted for in terms of sharing certain properties non-accidentally. I will take a closer look at the plausibility of content similarity and the role of non-accidental properties regarding concepts in chapters 4 and 5, respectively.

The property of *being a tomato* is what it is independent of what we take it to be;[38] and the presumption that there are different bodies of knowledge about tomatoes that are preferentially activated when we categorize that property in terms of what we take it to be does not support the claim that those bodies of knowledge *are* distinct co-referential concepts. In this respect, Machery's proposal could make a stronger case if it included a theory of reference that could help us understand how his distinct bodies of knowledge with little in common refer to things in the world. However, his proposal is not explicitly committed to any known theory of reference, and it offers no alternative one. Hence, the co-referential character of his distinct bodies of knowledge remains unexplained as well.

Problem 2: Distinct bodies of knowledge can't account for Publicity. The notion of concept publicity is based on our strong intuition that conceptual identity is somehow maintained between different people and across different time slices of the same individual. A good theory of concepts should explain, for example, how the concepts HOUSE, TONIGHT, etc. that someone other than me uses when they think that *Bernardo should be hosting the film group at his house tonight* are the same that I use when I think that *tonight might not be the best moment for a film night at my house*. The shortest way to show that Machery's theoretical proposal fails to explain how the distinct bodies of knowledge can be shared among different people and the same person across time is to establish that it fails because we cannot secure shared reference for Prototype, Exemplar, and Theory theories, as argued above (see Chapter 2, section 3.2). But Machery's view raises its own particular difficulties in this respect, given that the bodies of knowledge it posits are meant to be *distinct* default entities underwriting *distinct* cognitive processes (see 2.1 above).

Surely, someone might think this objection is question-begging since a characterisation of the publicity requirement in terms of conceptual identity takes it for granted that there are such things as concepts. Alternatively, if one does not think there are concepts, then one could claim that the reason why people who use, say, the same words manage to communicate fairly well with each other is that they are deploying similar, not identical, bodies of knowledge. However, such a counter-argument depends upon an acceptable notion of similarity, and it is unclear how any

[38] Or, for that matter, independent of the fact that we can think about it in different respects.

such notion is to be unpacked. As we will see in Chapters 4 and 5, while it is controversial that content similarity between different bodies of knowledge of the same type (e.g., two tokens of a BIRD-prototype) can guarantee publicity, it is probable untenable that similarity between different bodies of knowledge that are not of the same type (e.g., a token of a BIRD-prototype vs. a token of a BIRD-theory) can do so.

Given the dichotomy stated in Problem 1, I maintain there is one of two ways in which we can say that Machery's view fails to account for the publicity constraint. On the one hand, Machery's view will inherit the very problems of the Prototype, Exemplar and Theory theories to satisfy this constraint in case his three bodies of knowledge are literally identified with the particular complex mental representations posited by those three theories, respectively. On the other, if Machery's proposal of bodies of knowledge remains neutral about their representational nature, then this proposal faces the additional challenge of explaining how the same distinct bodies of knowledge can be shared and used in distinct processes both among different people and the same person across time.

Machery (2009) proposes the criterion of default knowledge for establishing what it is for a given person to have distinct concepts about a given category. He also claims that an account of how different people can share the same concept (or body of knowledge, for that matter) is scientifically irrelevant (2010). I think we should object both to Machery's claim that concept sharing is irrelevant for psychological explanation and to his criterion for establishing the possession of distinct kinds of concepts about any particular category. There are different reasons for this objection. I will focus on two: the compelling case of intentional generalizations and the implausibility of the default hypothesis.

Machery's Heterogeneity Hypothesis is not intended to subserve psychological explanation qua intentional explanation—i.e. one that appeals to mental states in order to explain people's behaviour. And it is not clear how it could help provide an alternative explanatory framework that subsumes the explanatory power of intentional explanations. He also avoids discussing views of concepts that take the representational status of concepts to be at the centre of psychological generalizations

(Rey 2009; Edward 2010). Neglecting the explanatory advantages of shared concepts in intentional generalizations is the price Machery has to pay for this omission.

One of the most prominent features of intentional explanations of behaviour is their generality, which depends on concept sharing (e.g., Prinz 2002). The assumption is that different people's actions can be explained as being motivated by the same propositional attitudes, typically beliefs and desires, provided that those attitudes are composed of the same concepts. For example, suppose Jack went to the farmers' market in Edinburgh because he desired fresh apples and believed that he could find some there. And suppose Eva went to the farmer's market in Barcelona for the same reasons. In these two cases, the same intentional explanation can subsume Jack and Eva because the attitudes motivating their actions are assumed to be composed of precisely the same concepts. In this sense, concept publicity is a requirement for intentional explanations to generalize. Hence, the intentional generalizations about peoples' behaviour that concept sharing enables is an explanatory advantage that any scientific understanding of the mind should be reluctant to easily dismiss.

Machery's view of distinct bodies of knowledge is inadequate to outperform the generality of intentional explanations (see Chapter 4). We all have different knowledge about any given category and our knowledge is empirically revisable. Likewise, it is also difficult to establish what particular knowledge about a category is collateral and what knowledge is defining of it. It is therefore difficult to establish how Machery's bodies of knowledge can be distinct or clearly demarcated from one another. The Heterogeneity Hypothesis proposes a solution for the problem of demarcation in terms of *default* knowledge, but there are reasons to think that this solution can't work.

"Default" is the hypothesis that some bodies of knowledge are preferentially (or automatically) retrieved from long term memory when one is categorizing, reasoning, drawing analogies, making inductions, and so on (e.g., Machery 2010, p. 196). Machery's view admits that the processes where these bodies of knowledge are involved can include non-default information when default information is not sufficient to solve a cognitive task. He calls "background knowledge" to any knowledge that is not constitutive of default bodies of knowledge. Nonetheless, the proposal is that only default knowledge can be said to be constitutive of any given

prototype, exemplar or theory. I disagree that "Default" offers a plausible way to distinguish what is constitutive of Machery's version of prototypes, exemplars and theories because the assumption that default and non-default knowledge can be clearly demarcated is untenable.

Machery grounds the notion of distinct default bodies of knowledge on the assumption that people's judgments under pressure are reliable indicators of their existence. However, cognitive tasks yield different results depending on whether or not time pressure is involved (e.g., Keil 1989; Rey, 2009, 2010; Blanchard 2010). Importantly, there are paradigmatic cases where the absence of time pressure shows that a theory approach to concepts does much better than similarity-based approaches in more or less reasoned categorization tasks, especially where the latter fail to consider hidden explanatory relations between features (e.g., Keil et al. 1999; Murphy 2002, c.6). Machery (e.g., 2010, section 3) is aware of this empirical evidence but considers it is still compatible with the role of background knowledge that is not contained in a given default body of knowledge. However, if this is so, then Machery's notion of theory as default knowledge would fail to account for the most interesting explanatory role of theories as advocated by Theory theorists (Jacobson 2010). This is the price Machery has to pay for grounding the theoretical independence of prototypes, exemplars and theories on the experimental criterion of time pressure.

Machery's view of distinct bodies of knowledge is committed to the hypothesis that these bodies of knowledge participate in similar yet independent cognitive processes. However, concept theorists often agree that different constructs may co-exist, and this is particularly the case of Theory theorists who think none of the current psychological theories is independently able to account for all cognitive phenomena (Murphy 2002, p. 64). In order to undermine this interdependence, Machery needs to provide stronger grounds for defending the existence of conceptual heterogeneity—and the related multiple independent processes—on the basis of Default. Given that the proposed bodies of knowledge can't be clearly demarcated, it is difficult to establish over what kind of shared bodies of knowledge psychological explanations can generalize.

Problem 3: Distinct bodies of knowledge can't account for Compositionality. Typically, a concept is said to compose in case compounds concepts are formed as a function of their constituent concepts and some rules of combination (Prinz 2002). A paradigmatic formulation of the compositional combination of concepts is in terms of contents. Suppose, for example, ROUND and TABLE are the constituents of the concept ROUND TABLE. The compositionality constraint states that it is impossible for an individual to possess the complex concept ROUND TABLE and not possess ROUND (or TABLE, for that matter). Thus, the meaning of the complex concept is a function of the meaning of its constituents. The shortest way to show that Machery's distinct bodies of knowledge fail to compose is to establish that they fail to do so because prototypes, exemplars and theories fail to do so as well (See Chapter 2, section 3.2.). But Machery's view raises additional difficulties in this respect.

The assertion that Machery's view of distinct bodies of knowledge inherits the problems that the Prototype, Exemplar and Theory theories have in accounting for concept compositionality only applies to the extent that Machery's constructs are literally identified with structured mental representations as per those three theories, respectively. However, as we know, Machery's bodies of knowledge can also be understood in terms of weaker identification to the traditional understanding of prototypes, exemplars and theories, namely, one that is neutral to their representational nature.

According to this alternative understanding, Machery's version of these three constructs reduces to an appeal to the kind of information that those three constructs are supposed to encode. Another way to characterize this information is in terms of default knowledge about different aspects of a given category stored in long term memory. The problem with this characterization is that it is non-committal (and, therefore, unrevealing) about the format of the inner states that serve to carry that information. In other words, what is missing is an account of what some (e.g., Bechtel 2008) have called the representational vehicles of our mental contents. So, to the extent that Machery's distinct bodies of knowledge are actually about categories in the world, i.e., to the extent that they can be said to have contents, we are owed an

account of the format of vehicles that play a role in the cognitive processes where the compositionality of mental representations is required.[39]

Not all kinds of representational formats can be said to compose. From Chapter 2, we know that prototypes, exemplars and theories are three notions of complex mental representations that can't satisfy the compositionality constraint. If you are neuroscientist and think that some kind of iconic (or map-like) representations of certain patterns of neural activity are a good model of our mental mechanisms, then you may also find it difficult to show how these representations are formed as a function of their constituent elements. The notion of distinct bodies of knowledge is particularly problematic to satisfy this constraint because it is silent about their representational format. In this case, the requirement is that (at least in some cases) a given distinct compound body of knowledge is formed as a function of its constituent bodies of knowledge plus some rules for combination. As a result of this, the content of the compound body of knowledge should be directly inherited from the contents of its constituent bodies of knowledge. However, we are not told how this could be accomplished.

Another possibility is that Machery's version of prototypes, exemplars and theories are capable of being combined with one another in order to account for the typical open ended set of concepts and thoughts we seem to be able to entertain. This strategy is inappropriate for, at least, two reasons. On the one hand, these allegedly co-referential bodies of knowledge tend to include information that can't easily be said to be constitutive of only any one of them (e.g., Virtel and Piccinini 2010; Murphy 2002, c. 3). On the other hand, prototypes, exemplars and theories are, by hypothesis, supposed to take place in distinct cognitive processes, so combining them in some processes underlying our cognitive capacities directly violates the Heterogeneity Hypothesis.

[39] As a case in point, our boundless capacity to entertain new and unique thoughts is typically assumed to required concept compositionality, and a good explanation for this compositionality is that format of the vehicles of thoughts (qua compound representations) inherit the format of the vehicles of its constituent representations.

3.2. The case of atomistic theories of concepts

Prototype, Exemplar and Theory theories are committed to the view that what makes a concept the very concept that it is *is* its relations to other concepts (see Chapter 2). Alternatively, atomistic theories are theories according to which what makes a concept the very concept that it is is how it is related to the world (Margolis 1998). Whereas prototypes, exemplars and theories are inherently vulnerable to the problems described in section 3.1., conceptual atomism is not.

I do not intend to provide an exhaustive account of atomistic theories of concepts. Instead, the purpose I have in mind is to support the claim that Machery's assumption that a good theory of concepts is unviable (because the Prototype, Exemplar and Theory theories have failed to provide one) is unwarranted. As I want to argue, the reason for this is that there is at least one plausible theory of concepts that does not identify concepts with prototypes, exemplars or theories. To accomplish this purpose, it will suffice to make this case by illustrating, in a very succinct way, the explanatory advantages of the Informational Atomism defended by Fodor (e.g., 1998).

To begin with, Fodor's Informational Atomism (IA, henceforth) holds that concepts have no structure and that the intentional content of an unstructured concept is its referent. The thesis that the content of a concept is exhausted by its referent should be understood in terms of the theory of reference that IA is committed to, namely, Fodor's (1990) Informational Semantics. Essentially, this theory holds that the content of a concept is determined by the concept's standing in an appropriate causal relation to the things in the world it represents (Laurence and Margolis 1999). Fodor conceives of this relation as a nomic relation between concepts and the things that fall under them. One common way to characterize this relation is to say that there is a counterfactual-supporting causal relation between the tokenings of a concept and the things in the world this concept expresses (Prinz 2002). Thus, for example, the reason why the concept FROG represent frogs is that there is a causal law controlling the informational connection between the property of *being a frog* and the tokenings of the concept FROG. This connection is informational because the concept FROG is said to "carry" information about the property in the world that it expresses (i.e., *being*

a frog). There is more to say about Fodor's theory of reference [40] but this characterization suffices my present purpose to show how IA accommodates the requirements discussed in section 3.1.

Basically, the intentionality requirement is satisfied insofar as the content of the unstructured (i.e., primitive) concept is determined by a reliable causal relation that the concept bears to and the property it represents. In this sense, IA does not only provide a naturalistic solution to the intentionality requirement, but it also avoids the problems of content stability other theories have. Theories that identify concepts with structured mental representations normally hold that concept identity is determined by the internal structural relations a concept bears to other concepts. The result is that the variability of the inferential relations in which the concept participates constrains the variability of the content of the concept. By contrast, IA holds that concept identity is determined by their intentional content. It is how a concept is related to the world that determines concept identity. Thus, this externalist mode of individuation accommodates the publicity requirement because it guarantees that different people and the same person across time share literally the same concept, independent of their culture and knowledge. People with different beliefs can share the same concepts insofar as those concepts stand in an appropriate nomic relation to the properties that cause them. Another way to illustrate the fact that people literally share the same concept is in terms of the type/token relation. According to IA, primitive concepts are presumed to satisfy such a relation in that people can have tokens of literally the same concept type.

The following assumption underlies Fodor's thesis that most lexical concepts have no internal structure: in order to account for the mental processes underlying human intelligent behaviour, it is not necessary to assume that inferences are constitutive of cognitive processes. All the theories characterized in Chapter 2 (section 2)[41] take it for granted that our inferential processes depend on the structure of concepts that those theories deem as essential to concepts. However, as Fodor suggests, there can be a more basic level at which it is possible to hypothesize that thought is compositional and that inferential processes are not incompatible with the

[40] Consider, for example, the way this theory responds to cases of an erroneous application of a concept (see, e.g., Fodor 1990; Laurence and Margolis 1999; Prinz 2002).
[41] Definitional, Prototype, Exemplar and Theory theories of concepts.

possibility of there being only unstructured concepts. This possibility requires accepting that complex thoughts (or mental representations)[42] can be formed as a function of a limited set of primitive lexical concepts and some rules for combination. Given that the content of primitive concepts is causally determined by a reliable relation to properties in the world, the content of complex mental representations can be said to be directly inherited from the content of its constituents.

3.3. The Heterogeneity-Eliminativism Fallacy

I am now in a position to show why the eliminativist conclusion about concepts does not follow from the distinctness of prototypes, exemplars, and theories, considered as putative candidates for the psychological role of concepts. There are two general reasons why we should hesitate before drawing such an eliminativist conclusion.[43] One is that there may be a scientifically valuable superordinate kind, which includes as members a number of distinct sub-kinds. The other is that we should not be confident that the list of prototypes, exemplars, and theories exhausts the possible candidates for the role of psychological concepts.

Specifically, in relation to the second of these general points we can see that the following argument bearing upon the Heterogeneity Hypothesis does not follow from its premises:

---

[42] IA admits that there are rules for molecule formation that can account for the psychological fact that there are complex mental representational structures, but only on the basis of the individuation of their constituents. This does not necessarily imply that complex mental representations (e.g., thoughts) are to be identified with concepts, pretty much like, e.g., a brick wall is not to be identified with a big complex brick. IA holds that concepts are primitives and it would be paradoxical to accept that *bigger unstructured concepts* are formed out of *smaller unstructured concepts*. Note that this claim does not rule out the possibility that certain structured mental representations (e.g., prototypes) can be the referents of unstructured concepts (e.g., an atomic concept whose referent is a BIRD-prototype). Complex mental representations are things in the word, so concepts can represent mental representations that are not concepts (e.g. prototypes as computationally-based representational structures that encode background, non-constitutive knowledge about the content of other—complex or simple—mental representations).

[43] In fact there are at least three reasons why the heterogeneity of supposed members of a kind does not establish the non-existence of such kinds. For there can also be highly abstract kinds, instances of which are known to be heterogeneous, even though there may be important nomic generalizations involving such kinds. Consider, for example, the kinds *predator* and *having mass of 1 kilogram*. However, I here mention just the first two reasons, since I am not claiming that concepts are kinds of such an abstract generality as the kinds of general physics.

Premise 1:	If a theoretical notion does not pick out a single homogeneous natural kind, we should eliminate it from scientific taxonomy
Premise 2:	The psychological notion of concept is used to refer to prototypes, exemplars and theories
Premise 3:	Prototypes, exemplars and theories do not form a single homogeneous natural kind altogether
Conclusion:	We should eliminate the notion of concept from the scientific taxonomy of psychology

As stated at the beginning of section 2.1 above, the first step of Machery's argumentative strategy for supporting the Heterogeneity Hypothesis involved establishing that for each object, substance, event, etc., there can be several concepts. For Machery (2009, 2010), this undertaking reduces to showing that, if the class of concepts includes prototypes, exemplars and theories, then "concepts divides into kinds that have little in common" (p.77). However, this preliminary requirement presupposes an unwarranted assumption, according to which an adequate unifying theory of concepts must identify concepts with some kind of complex mental representations, notably, Prototypes, Exemplars or Theories. This assumption is unwarranted because it overlooks the possibility of a plausible theory of concepts that is not committed to the view that concepts are structured mental representations. I have claimed that Fodor's IA is that theory, but Fodor's theory is not the only available atomistic option (see, e.g., Schneider 2011).

Though IA is not without problems, it provides a plausible solution to at least three important requirements normally imposed upon a good theory of concept, namely, intentionality, publicity and compositionality. As discussed in sections 3.1 and 3.2, Machery's view of concepts not only fails to satisfy them all, but it also has further related problems. Moreover, in characterizing IA, I have also provided grounds to support the thesis that structured concepts might be both explanatory and metaphysically unnecessary.

I agree with Machery that the Prototype, Exemplar and Theory theories have failed to provide a unifying theory of concepts, but I disagree that there can be no unifying theory of concepts at all *because* those three theories have failed to provide one. The unwarranted assumption mentioned above allows us to unveil a hidden premise:

Premise 4: Either the class of concepts includes a single type of complex mental representation, *or else* it includes prototypes, exemplars and theories qua *distinct* bodies of knowledge.

This new premise makes explicit the false dilemma contained in the argument above. For Machery's Heterogeneity Hypothesis to be plausible, one should have to accept Premise 4 or a similar one. However, since there is, at least, one plausible theory of concepts that does not identify concepts with complex mental representations, one does not have to accept Premise 4. Indeed, consider that one can agree that this premise contains an unwarranted disjunction even if one is not an IA advocate, since there is even, at least, one plausible theory of concepts qua structured mental representations that does not need to identify concepts with prototypes, exemplars and theories (e.g., Laurence and Margolis 1999).

As we saw in Chapter 1, theoretical notions are not to be abandoned from the vocabulary of a given science simply because they fail to clearly pick out a putative natural kind. Many highly vague theoretical concepts can play a fruitful role in explanatory adequacy and some of these concepts have proved to be crucial for enquiry to proceed (e.g., Cowie 2009). So, Premise 1 is false, unless other supporting premises are supplied. Machery argues that the notion of concept is also explanatory idle and that psychologists should focus on prototypes, exemplars and theories. However, Machery has not made the case that those three psychological structures can do the job normally attributed to concepts.

Given that Premise 1 and Premise 4 are false (unless other supporting premises are supplied), nothing really hangs on Premise 2 and Premise 3 being true or false, even if prototypes, exemplars and theories were distinct natural kinds (since being a natural kind and being a concept are not necessarily co-extensive) and even if IA was false.

4. Conclusion

In this chapter, I took issue with a recent eliminativist thesis regarding concepts put forward by Edouard Machery. In order to fulfil this purpose, I first

addressed Machery's main aspects of his argument for concept eliminativism. This involved an account of the main tenets of his Heterogeneity Hypothesis and the general argumentative strategy for supporting these tenets.

After that, I discussed some familiar problems connected with the psychological theories of concepts Machery considers relevant for his own view. On the basis of that discussion, I argued that Machery's theoretical framework for conceptual heterogeneity not only inherits the same problems of the mentioned theories, but also introduces new problems that render his eliminativist proposal less viable.

Finally, I presented and discussed a version of Machery's general argument bearing upon the Heterogeneity Hypothesis. I argued that it contained an unwarranted assumption, according to which an adequate unifying theory of concepts must identify concepts with some kind of complex mental representations. To support my point, I went over a succinct account of Fodor's Informational Atomism and argued that the eliminativist conclusion of the argument does not follow from the premises, even if Informational Atomism was false. Hence, the conclusion is that Machery's argument fails and that we should therefore reject his thesis about the elimination of concepts from the theoretical jargon of cognitive science.

Chapter 4

Content and Bodies of Knowledge

1. Introduction

According to the type of concept eliminativism that Machery (2009, 2010) advocates, cognitive scientists should avoid the notion of concepts for two reasons: first, psychologists are wrong when they assume that concepts amount to a natural kind; second, keeping this notion as part of the theoretical vocabulary of psychology prevents scientific progress. His view is that the class of concepts is not really a homogeneous class, but a collection of at least three distinct yet co-referential *bodies of knowledge* (henceforth, BoKs) characteristically used in independent cognitive processes.

One of the ways in which this type of eliminativism can be challenged is by arguing that some relevant reconstruction of the general argument against concepts is committed to certain important unwarranted assumptions. That was the aim of Chapter 3. Another way of challenging concept eliminativism of this sort is by arguing that, in some important respect, the costs of eliminating the relevant theoretical term outweigh the benefits of keeping it. This is the aim of the current chapter.

More specifically, the aim of this chapter is twofold. I will first argue that Machery's view of BoKs can't avoid the problem of content, i.e. providing an answer to the question of how it is that these putative BoKs can attain their intentional content. Then, I will show that Machery's proposal that psychologists should replace the notion of concept with his BoKs is a bad proposal because there is no plausible theory of content in sight that can be made to work for these BoKs.[44]

[44] I am indebted to discussions and helpful comments by Bernardo Aguilera about topic of this chapter, with whom I am co-authoring a related manuscript currently under peer review.

2. What the problem is

Many cognitive scientists and philosophers agree that concepts are the basic constituents of thought and, therefore, that they must play a central role in psychological explanation. Scientists' generalisations and predictions about our conceptual abilities are based on the assumption that it is concepts that allow us to coordinate our behaviour with respect to the features of the environment that concepts represent. In this sense, psychological theories of concepts are expected to explain many different yet interrelated phenomena. Among the most prominent phenomena are concept acquisition, categorization, inference, and the stability of meaning. As we saw in Chapter 2, this latter phenomenon is usually understood in terms of the capability of concepts for being shared by different people—or the same person at different time slices—when deploying thought processes about the same referent.

Another reason for the centrality of concepts in the understanding of our cognitive capacities is that they can help explain the property that our thoughts have to represent objects (individuals, relations, etc.) in the world. Theorists call this phenomenon the 'intentionality of thought'. With respect to this phenomenon, one can roughly characterise the relevant role of concepts in the following way. Contemporary research in the human mind has traditionally identified concepts with mental representations and, given the representational capacity of concepts, researchers have taken it for granted that the intentionality of thoughts reduces to the intentionality of their constituent concepts. In turn, the way the conceptual capacity of representing is explained is by appealing to the notion of *content*, a notoriously divisive notion that has been the subject matter of extensive debate and theorising (e.g., Putnam 1975; Searle 1983; Dretske 1988; Millikan 1989; Fodor 1990). Regardless of the variety of views on this notion and its role in an account of the intentionality of thoughts, it is relatively uncontroversial to say that concepts are typically understood as content-bearing and that an explanation of content is deemed a desideratum for an adequate theory of concepts (Fodor 1998; Prinz 2002).[45]

[45] Importantly, content is not only relevant for an account of how we manage to interact with external entities, but it is also essential for an account of the semantics of language. Influenced by Grice (1957), many philosophers maintain that linguistic expressions have meaning by virtue of deriving their contents from the content of the concepts that those expressions are used to express.

Despite this important role that concepts customarily play in psychological explanation, the very scientific relevance of the notion of concept has, as we have seen, been called into question by Machery (2009). Machery claims that this notion should be displaced altogether from the theoretical vocabulary of psychology. His alternative suggestion is that different types of *bodies of knowledge* (BoKs) can do the job of explaining the same cognitive phenomena that those researching into the psychology of concepts have tried to account for by appealing to the theoretical notion of concept. As critically examined in Chapter 3, Machery's argument is that these BoKs are so distinct from one another that they cannot be subsumed by a single general notion of concept, from which he concludes that this notion really fails to pick out a genuine psychological kind.

In this chapter, my aim is to critically examine Machery's notion of BoKs with respect to the problem of content. Theories of concepts within scientific psychology have traditionally provided us with possible accounts of mental content, so it is perfectly legitimate to ask whether Machery's alternative to concepts can be supplemented with an account of content. Until the time of writing the present thesis, Machery has not provided us with such an account and I will attempt to show that no plausible notion of content can be combined with his heterogeneous BoKs. If this is right, then a psychology based on BoKs will be quite unable to account for the intentionality of the representational mind, clearly a very serious flaw. My strategy will be the following. In section 3, I will revisit Machery's case for the elimination of concepts and a characterisation of his proposed alternative to concepts. The purpose of section 4 is to state Machery's position regarding the problem of content and why this is a problem his view can't avoid. Section 5 will introduce two broad approaches for characterising the relation between concepts and what concepts are about. The most prominent views of content determination will be described with respect to these two approaches. In section 6, I will assess the possible application of the views presented in the previous section to Machery's view of BoKs. My conclusion will be that none of those accounts of content can be adapted to Machery's alternative to concepts, which points to bleak prospects for his case for concept eliminativism.

3. Machery's distinct bodies of knowledge

One of the questions Machery (2009) asks, in connection with the current empirical study of concepts,[46] is whether or not we have good reasons to think that the notion of concept psychologists make use of really designates a homogeneous class. If the answer to this question is positive, then concept theorists might be justified in inferring that such a notion is likely to pick out a natural kind and, therefore, that there must be a correct theory of concepts—probably, a version of one of the revisionist theories discussed in Chapter 2 (section 2.2) which attempt to displace the classical theory of definitional concepts. If the answer is negative, the idea of a single correct theory of concepts could be put into question.

The assumption that, if one of the available psychological theories of concepts is true, then none of the others is true is what Machery calls "the Received View." An interesting reaction to this view is Machery's thesis that the class of concepts really divides into several heterogeneous kinds, a view that he dubs the *Heterogeneity Hypothesis*. He takes these kinds to be distinct BoKs. Roughly, a BoK is said to be information about a certain category that is preferentially retrieved from long-term memory to be used in cognitive processes underlying our higher cognitive competences. It follows from this characterisation of a default BoK that not all knowledge about a given category can count as a relevant BoK for solving a particular task. Indeed, Machery distinguishes between *default knowledge* (namely, knowledge about a given category that is preferentially retrieved) and *background knowledge* (namely, knowledge about the same category that isn't preferentially retrieved). Importantly, while background knowledge is said to be context-dependent, default knowledge is said to be context-independent because it can be retrieved in all contexts (Machery 2015). These two types of knowledge do not have clear-cut boundaries, since, according to Machery, people's default BoKs about a given category do not need to be exactly the same for them to count as BoKs of the same type and about the same category. He supports this claim by resorting to the view of natural kinds developed by Boyd (1991, 1999). According to this view, a natural kind is a class of things that form a *Homeostatic Property Cluster* (HPC), i.e. a class whose members

[46] For a general evaluation of main theories of concepts, see Laurence and Margolis 1999, Prinz 2002, c. 3-4 and Machery 2010, c. 4. For a detailed account of findings in experimental cognitive psychology, see Murphy 2002.

share some cluster of properties in virtue of certain causal mechanisms. These mechanisms sustain a cluster of properties which are jointly sufficient (but individually need not be necessary) for determining class membership.

Machery (2009)[47] examines the models of concepts and cognitive processes proposed by the main psychological theories (i.e., Prototype, Exemplar and Theory theories) and concludes that these models have little in common with one another, "given the properties that are relevant to characterize concepts" (p. 5). His suggestion is that prototypes, exemplars and theories, if they exist, must amount to distinct BoKs because they do not form a single natural kind with many common properties, as Boyd's notion of HPC requires. In his view, this realisation is sufficient reason for displacing the term 'concept' from psychological taxonomies, since it is common methodological practice for any given science to eliminate from their classificatory schemes those theoretical terms that do not pick out natural kinds. Thus, given the assumption that the term 'concept' does not pick out a natural kind which is able to support many inductive generalisations, Machery defends the view that the existence of his proposed distinct BoKs provides pragmatic grounds for the elimination of the theoretical notion of concept.

Machery's proposal is essentially based on the view that the theories that identify concepts with prototypes, exemplars and theories are not rival theories competing to provide an account of the same phenomenon because they really provide adequate descriptions of different varieties of categorisations (Hill 2010). Thus, in line with the Heterogeneity Hypothesis, his contention is that what psychologists have taken to be three different models of concepts should be better understood as three different BoKs playing a central role in distinct and independent cognitive processes. Hence, contrary to the Received View, the explanatory advantages of any one of these theories would not yield evidence against any of the others. According to the alternative view, for each category (i.e., a certain class of things), an individual typically has three fundamental concepts, which Machery conceives of as distinct co-referential BoKs.

How can we tell when someone is using distinct BoKs which refer to the same category? Well, to answer this question Machery offers two individuating criteria,

[47] See also Machery (2005 and 2010).

‘connection’ and ‘coordination’, and establishes the following condition: whenever two elements of information about x, A and B, fulfil either of these criteria, those elements of information can be said to belong to two distinct concepts. Machery (2010) characterised the connection and coordination criteria in the following way, respectively:

> 1. If retrieving A (e.g., water is typically transparent) from long-term memory and using it in a cognitive process (e.g., a categorization process) does not facilitate the retrieval of B (water is made of molecules of H2O) from long-term memory and its use in some cognitive process, then A and B belong to two distinct concepts (WATER1 and WATER2).

> 2. If A and B yield conflicting judgments (e.g., the judgment that some liquid is water and the judgment that this very liquid is not water) and if I do not view either judgment as defeasible in light of the other judgment (i.e., if I hold both judgments to be equally authoritative), then A and B belong to two distinct concepts (WATER1 and WATER2). (p. 196)

Thus, while the first criterion allows for WATER-prototype and WATER-theory to count as two distinct BoKs, the second criterion is consistent with the possibility that two contradictory judgments (e.g., *penguins are birds* and *penguins are not birds*) are caused by different BoKs (e.g., PENGUIN-theory and PENGUIN-prototype, respectively), provided that one of them is defeasible.

According to Machery, the empirical fact that people’s categorisation judgments can be explained in terms of different BoKs is evidence that people do not make use of a single kind of concept and that, instead, we typically use different independent representations (notably, a prototype, a set of exemplars and a theory) for any given category. He rejects the less radical proposal that there might be just one kind of representation the components of which include different types of structures as the proponents of Hybrid Theories of concepts have claimed. As he goes on to argue, the reason for his rejection is that those theories have failed to explain what it means for those different BoKs to be parts of the same concept.

Machery’s general argumentative strategy moves from the claim that prototypes, exemplars and theories amount to distinct BoKs to the argument that those

three BoKs have very few properties in common because they really pick out different natural kinds (i.e., different HPCs) relevant for different and independent cognitive processes. More precisely, his claim is that prototypes, exemplars and theories are three different psychological entities that individuals retrieve by default from long-term memory to be used in cognitive processes underwriting our higher cognitive competences. If this was correct, as the evidence he reviews seems to suggest, then there would be grounds to contend that what psychologists attempt to explain by appealing to the notion of concepts cannot be accounted for by a single general scientifically relevant kind. Indeed, Machery's radical proposal is talk of concepts should be replaced with talk of at least three distinct yet co-referential BoKs. As Machery points out, his proposal is completely empirical and able to be revised by, for instance, future discoveries in neuroscience regarding concept location and neural dissociation that challenges his argument for heterogeneity. Similarly, Machery's revisionary proposal regarding the study of concepts is open to the possibility that the number of BoKs individuals have about a given category is not exhausted by prototypes, exemplars and theories.

4. What Machery says on the problem of content and why he can't avoid it

Machery (2009) points out that his usage of the term "knowledge" corresponds to that of psychologists', which he characterises as "any contentful state that can be used in cognitive processes" (p.8). It would appear that by adopting psychologists' standpoint regarding the study of cognition Machery's own view of BoKs need not be committed to providing an account of how it is that these admittedly contentful mental states attain their contents. The reason for this is double: one the one hand, psychologists are typically interested in characterising the type of knowledge that might be constitutive of contentful BoKs (relevant for some cognitive task), without specifically providing an account of how the contents of those mental structures are secured. On the other hand, since Machery's view of BoKs is a revisionist proposal that intends to question the viability of a scientific theory of concepts, it might seem that by challenging this latter project defenders of a revisionist view of BoKs are no longer expected to supply a suitable theory of content as one would expect concept theorists to do. However, as I will intend to show,

neither psychologists' typical explanatory interests in the study of cognition or Machery's revisionist proposal is a good reason to avoid the problem of content. Let us consider each of these issues in turn.

With respect to the first issue (i.e., psychologists' explanatory interests), Machery relies on his methodological contention that theorising about concepts in psychology and philosophy pursue different agendas. According to him, whereas psychologists are primarily interested in the role concepts may play in processes such as reasoning, learning and categorising, philosophers are mainly concerned with an account of the semantic nature of our thoughts. One way to describe this divergence of interests is in terms of a division of labour according to which, while psychologists' business is to explain people's behaviour by appealing to contentful states, philosophers are expected to provide a foundational account of the semantic properties of those mental states. Indeed, even though this is a view some philosophers (e.g., Fodor 1994; Peacocke 1992; Rey 1998) would gladly welcome, it is one Machery thinks we should treat with caution due to epistemological reasons. What he specifically warns us about is philosophers' development of theories of content on the basis of intuitions about the conditions under which one would find it plausible to ascribe to people certain attitudes with propositional contents. Machery supports this claim with evidence suggesting that people's intuitions regarding the ascription of propositional attitudes are neither uniform (e.g., Hewson, 1994) nor free from cultural variation (e.g., Nisbett et al. 2001; Machery et al. 2004). I agree with Machery that, even though the philosophical task of accounting for how it is that we can have contentful states should be taken with caution, there is no reason to think that this project can't be successfully developed. In that spirit, I will later on intend to explore what available theory of content could possibly be combined with Machery's theory of BoKs.

With respect to the second issue (i.e., the possible irrelevance of a theory of content), Machery betrays a hasty generalisation one could make explicit in the following way. The inference Machery seems to take for granted is that, if the justification of a theory of content is that it is expected to supplement a theory of concepts, then, as someone might say, it follows from the fact that there are no concepts that there is no justification for a theory of content. Hence, we should not expect that someone who is committed to the existence of distinct BoKs (and, hence,

to the non-existence of concepts) is committed to a theory of content that works for those BoKs. However, this inference is invalid because it wrongly takes it for granted that a theory of content is only relevant for mental representations we identify with concepts. Indeed, one does not have to go very far to find a relevant counter-example. Machery (2010) reduces an account of content to an account of a "semantic theory for our propositional attitudes" (p. 47), but this reduction is clearly at odds with his own view that knowledge does not keep to explicit propositional states. According to Machery, the kind of knowledge that is relevant for his proposed BoKs can also be "implicit or explicit [as well as] propositional, imagistic, or procedural" (p.8). Thus, what this latter characterization of knowledge suggests is that someone who defends the scientific relevance of prototypes, exemplars and theories qua distinct BoKs is not *ipso facto* free from an account of how those BoKs attain their contents.

Notwithstanding this latter consideration, Machery sometimes appears to incline towards a more radical motivation for ignoring issues connected with the problem of content in his proposed theoretical framework. If it is possible to make the case that the notion of concepts is theoretically idle—as someone might claim—, it might well be possible to make the case that the notion of content is theoretically idle too. Indeed, some think the idea that reference be deemed as a desideratum for an empirical theory of concepts is at least methodologically controversial (e.g., Machery et al. 2004; Piccinini and Scott. 2006).

However, opting for this alternative would be moving too fast for at least two related reasons. First, Machery's proposed BoKs are intended to replace concepts in the explanation of psychological phenomena formerly attributed to concepts. If this remark is correct, then those BoKs are expected to be involved in psychological processes whereby the perceptual input is classified into a collection of meaningful categories in virtue of which the world is made intelligible to us. Second, for the proposed BoKs to be able to do this job it is required that they do not only possess processing properties that help to account for *how* it is that these BoKs are used to discriminate and categorise, but also semantic properties that help to account for *what* it is that they discriminate and categorise. In other words, as is the case with research on concepts, research on distinct BoKs should be interested not only in the investigation of processing properties prototypes, exemplars and theories are expected to possess, but also in the investigation of the semantic properties that enable them to

refer to external objects in the world. This latter issue is particularly relevant for research on Machery's BoKs because different BoKs about the same category are assumed to be co-referential. Thus, for example, while it is important for the proposed theoretical alternative to concepts to able to establish the way in which three distinct BoKs about the category of cats (e.g., a CAT-prototype, CAT-exemplar and a CAT-theory) are used to categorise cats, it is also important to establish how it is that each of those BoKs can refer to external objects in the first place, viz., how it is that they can have the contents that they are supposed to have.

Machery refrains from providing a possible account of content for his distinct BoKs on the grounds that such an account is not expected to be part of the explananda of a psychological theory of concepts, and hence, of the explananda of a theory of BoKs. His view is that philosophical and psychological theories of concepts are two different projects because philosophers and psychologists working in concepts do not talk about the same thing when they use the theoretical term 'concept'. However, Machery (2010) has responded to his critics by acknowledging the importance of establishing how it is that prototypes, exemplars and theories can co-refer (see, p. 234). I agree with Machery in this respect and I also concur with his diagnosis that, when theorising about concepts, philosophers, not psychologists, have typically focused on the problem of content. But I disagree with both his contention that concept theorists in philosophy and psychology are not really talking about the same type of thing and the view that an account of content may not be relevant for a theory of concepts (or, *mutatis mutandis*, for a theory of BoKs).

I maintain that, just as in the case of concepts, an account of how it is that prototypes, exemplars and theories attain their contents should be part of the explanatory desiderata a complete theory of distinct BoKs should ideally explain. Indeed, such an account invites important interaction between concept theories in philosophy and psychology. In what follows, I will intend to show that a failure to combine Machery's BoKs with current theories of contents constitutes a reason to resist both the plausibility of BoKs as an alternative to concepts and Machery's case for concepts eliminativism.

5. Two broad approaches to conceptual content

Theorists have proposed different ways of characterising the relation between concepts and their contents, viz., the things in the world they are supposed to be about. Here, I propose to consider an initial distinction between two broad approaches to this relation: *Intensionalism* and *Referentialism*.[48] Intentionalism states that the content of a concept determines its extension. Referentialism states that the extension of a concept determines its content.

As I will broadly characterise it here, Intentionalism comprises two views of content determination. According to the first view, content is determined by 'Fregean senses' and, according to the second, content is determined by 'conceptual role'. In turn, Referentialism will be characterised in terms of its most prominent version according to which the conceptual content is fixed by some informational-cum-nomological relation to its referent, such as in the case of *Informational Semantics*.

With this distinction in hand, the aim of the following section will be an assessment of Machery's view of BoKs with respect to each of the two approaches to content introduced above. In this sense, the hypothesis I want to test is that, if it was possible to establish that Machery's view of distinct BoKs can be combined with either Intensionalism or Referentialism, his view would count as a serious alternative to concepts and his case for concept eliminativism would be strengthened.

6. Is there a viable account of content determination for Machery's proposed alternative to concepts?

Intensionalism and BoKs. I will begin by assessing Machery's distinct BoKs with respect to what I have called Intensionalism. Regarding this first approach, I broadly distinguished two main views of content determination (i.e., one based on Fregean senses and the other based on conceptual role) and I will now examine each of them in turn.

[48] This distinction is motivated, though not strictly in accordance, with the relevant distinction proposed by Fodor and Pylyshyn 2015.

Philosopher Christopher Peacocke (e.g., 1992) has popularised a neo-Fregean view of concepts that exploits Frege's distinction between *sense* and *reference* and the doctrine that linguistic expressions can have both of them, such that the sense of an expression contains the *mode of presentation* of its referent. Thus, for example, the expressions "giraffes" and "long-necked ruminants" can be said to express two modes of presentation (i.e., senses) of one and the same referent in the world. According to Frege, our access to the referent of an expression depended on whether we were able to grasp the sense of the expression. Peacocke proposes a theory that intends to legitimate Frege's ontology of abstract objects, identifying concepts with Fregean senses,[49] which are also said to be the constituents of propositions. Here, the central idea is a concept (i.e., a Fregean sense containing a mode of presentation of a given referent) is individuated in terms of its possession conditions. Thus, since Fregean senses are concepts and concepts are supposed to be the primary bearers of content, when a thinker satisfies the conditions to possess a Fregean sense, this sense (i.e., concept) ipso facto has content.

According to Machery, prototypes, exemplars and theories form three distinct psychological kinds because they are retrieved by default in order to be used in independent yet similar cognitive processes underwriting our higher cognitive competences. In this sense, something like an intensionalist neo-Fregean account of content would appear to be compatible with Machery's view of BoKs. Consider, for example, that a DOG-exemplar and DOG-prototype are distinguished from one another in that they are preferentially used in independent epistemic mechanisms, such as, say, exemplar-based categorisation and prototype-based categorisation, respectively. These mechanisms, as someone might say, could constitute conditions for the possession of three different modes of presentation of the same referent, which, in Machery's terminology, would be equivalent to three distinct BoKs about the same category. To make this claim more appealing, one could even argue that, when using a BoK an individual is really grasping a Fregean sense, since entertaining a certain BoK involves satisfying the corresponding possession conditions for entertaining a certain modes of presentation of a given referent.

[49] The proposal is to supply Frege's notion with a strategy of legitimation based on the conditions for the empirical application of senses, namely, the concept's *possession conditions* (see, e.g. Peacocke 1996).

Presumably, if this account of content could be made to work for Machery's BoKs, this strategy would offer many advantages associated with the neo-Fregean notion of content. First, if it is true that senses are what fix reference, this strategy would provide a solution to account for the intentionality of BoKs. Second, it would also explain how DOG-prototypes, DOG-exemplars and DOG-theories could be distinct cognitive mechanisms that share the same referent. The idea here is that senses containing different modes of presentation of a given referent could be conceived of as distinct co-referential BoKs, in a way that is analogous to how H2O and WATER in the propositions "water is liquid" and "H2O is liquid" are distinguished by means of grasping two senses of a single referent (water). And finally, this strategy would account for the stability of content and shareability of BoKs, since senses can be grasped at different times and by different individuals.

This move might appear difficult to accept for someone who thinks that neo-Fregean approaches have had a hard time trying to account for the ontological status of senses and, hence, explaining how they can be grasped by our representational mechanisms. However, the notion of senses can be combined with theories of concepts that do not hold the Fregean commitment that senses are non-mental abstract objects. For example, Prinz (2002) and Margolis and Laurence (2007) have suggested that the notion of senses could be adapted to a representationalist approach to concepts by applying the type/token distinction, where concept types are conceived of as some sort of abstractions and concept tokens as their particular mental realisations. Whereas Prinz highlights that the putative abstract ontological status of senses can be reconciled with the abstract status of mental-representation types, Margolis and Laurence have put forward a 'Mixed View' of concepts which accommodates senses as the putative meanings that mental-representations tokens can have or express. In any case, viewing senses in terms of a type/token distinction regarding concepts would be no more problematic than presupposing a metaphysical type/token distinction regarding properties in general, as it would be, for instance, to say that brown objects instantiate the property of *being brown*.

However, I believe that this neo-Fregean approach to content does not work for Machery's purposes. Note that, if our use of BoKs satisfies the conditions for grasping senses, then the existence of concepts would not be challenged by the fact that different BoKs are preferentially retrieved to be used in independent cognitive

mechanisms. It may be recalled from our initial characterisation that grasping a Fregean sense requires that a thinker satisfies the conditions that account for mastering a concept. Hence, insofar as those possession conditions take the form of psychological mechanisms that are capable of grasping a Fregean sense (or instantiating a concept type, for that matter), the relevant BoKs can be said to possess conceptual content. If this is correct, then there is no clear motivation for eliminating concepts, and, hence, for discussing distinct kinds of BoKs as a replacement for the notion of concepts.

Someone might want to insist on combining this neo-Fregean approach to BoKs and claim that equating BoKs with Fregean senses can ascribe content to BoKs without having to identify senses with concepts. If the notions of sense and concept were thus dissociated from one another, then the fact that those BoKs can capture Fregean senses would not threaten Machery's case for concept eliminativism. However, this would imply pruning the notion of sense to the point of losing its significance. Consider that characterising content in terms of Fregean senses brings along features which have traditionally been associated with the conceptual, such as compositionality, reference determinacy and force independence. Indeed, resorting to this Fregean approach is a standard strategy used for establishing criteria for concept possession (Gunther 2003), the satisfaction of which some think allows for concept individuation (e.g., Peacocke, 1992, 1996; Cussins 1990, 1993). Thus, my conclusion is that an intensionalist neo-Fregean view of BoKs is incompatible with Machery's case for concept eliminativism.

As I pointed out at the beginning of this section, Intensionalism can also be associated with the view that content is determined by conceptual role. According to this view, the content of a concept is determined by (at least some of) its inferential connections to other concepts. Standard versions of the Prototype, Exemplar and Theory theories of concepts are good cases in point. According to these theories, a concept is some kind of mental representation that bears a particular type of structural relation to other concepts. Thus, whereas the content of a prototype concept is determined by a set of relations of statistical frequency to its constituent concepts, the content of an exemplar concept is determined by its similarity relations to the representation of its exemplars as stored in memory. Likewise, the content of a theory

concept is supposed to be determined by its inherent inferential relations with other concepts within a given theory (or belief system, for that matter).

However, these theories share a common problem connected with the very inferential character of content determination they put forward. Given this character, it becomes evident that, because the corresponding inferential relations are subject to inevitable variation, the contents of the proposed models of concept are inherently unstable.[50] For example, if an individual's prototypical violin changes from wooden-made to made-of-diverse-materials, then, according to Intensionalism based on conceptual roles, her old and new prototypical violins can be said to be BoKs with different semantic structures and, hence, different contents. Similarly, people whose VIOLIN-prototypes differ in some respect would not be able to think or talk about the same thing when using their VIOLIN-prototypes. What this example suggests is that BoKs with unstable inferential structures both within and across individuals are unsuited to support generalisations about various phenomena (e.g., how people reason, learn and communicate with one another) in virtue of BoKs that refer to the same category. Put differently, if Machery's proposed BoKs are to attain their contents from their inferential relations to other prototypes, exemplars and theories, then this account of content determination is bound to inherit the *problem of content stability* stemming from the models of conceptual representation these BoKs are based on.

Admittedly, the way I have characterised the problem of the stability of content determined by conceptual role is based on the assumption that conceptual content is constituted by every inference that a concept is involved in. This is what Fodor and Lepore (1992) have described as semantic holism. Someone might claim, however, that holism is not a challenge to content determination by means of conceptual role because a limited number of inferences, instead of every one of them, can do the job of providing our mental states with stable content. This view is usually committed to the idea that content similarity, not content identity, is sufficient for the stability of meaning required for different people (or the same person over time) to be said to share the same concepts (e.g., Harman, 1993). Regarding this alternative, Fodor and Lepore (1992) have replied that constancy of content fails to be accounted

[50] See Fodor and Lepore 1992.

for in terms of a limited number of inferences without presupposing some sort of identity between inferences that overlap within or between individuals. A related problem to which Fodor and Lepore draw our attention is that, even if we granted that there is some invariant set of inferences that is constitutive of content, such an account of content would have to appeal to some kind of analytic/synthetic distinction in order to specify that set of inferences. The problem arises because, according to these authors, Quine's critique of this distinction is widely thought to have shown the notion of analyticity to be untenable, at least for employment in fundamental philosophical theorising.

In response to Fodor and Lepore, Greenberg and Harman (2006) argue for an ecumenical version of Conceptual Role Semantics (CRS) which they think is immune to the problems we have just mentioned. In this version of CRS, conceptual roles are identified with several *aspects of use* of mental representations in thought. These aspects of use do not need to be restricted to internal inferential relations, which, according to Greenberg and Harman, correspond to just one of the several possible aspects of conceptual role (i.e., one of the several ways mental representations are used in thought).[51] In these authors' view, whereas conceptual roles are said to be part of what determines that a mental representation has a given content, they should not be deemed constitutive parts of the content of the mental representation. Hence, as Greenberg and Harman claim, CRS is committed to a notion of content stability that is based on content similarity, without presupposing sameness of content. This is a theory of content determination deserves our attention since it could provide Machery's view of co-referential BoKs with a plausible notion of content stability by means of conceptual role. However, I think there is some important incompatibility between Greenberg and Harman's notion of content similarity determined by conceptual role and Machery's putative co-referential BoKs.

According to Greenberg and Harman, the content of a mental representation derives from use. If this view was to work for Machery's BoKs, then what a given BoK represents would have to be determined by the roles of the BoK in a person's cognitive economy. Thus, for example, what a prototype (e.g., BIRD-prototype)

[51] According to these authors, other aspects of use may include perceptual representation, recognition of implications, modelling, inference, labelling, categorization, theorizing, planning, and control of action.

represents is determined by the way it is used to categorise certain things in the world (viz., birds), and the same applies for exemplars and theories, mutatis mutandis. Moreover, given the assumption that content similarity is a sufficient condition for content stability, the similar ways different people may represent a given object, say, prototypically could provide grounds for those people to be said to possess co-referential prototypical representations of that object. And the same can be said about the ways different people may represent a given object by means of exemplars and theories, respectively. However, even if we granted that the similarity between two tokens of the same type of BoK can be enough to explain their co-referentiality (e.g., that two BIRD-prototypes are both about birds), things become fuzzier when it comes to different types of BoKs which are also supposed to be co-referential (i.e., BIRD-prototypes, BIRD-exemplars and BIRD-theories). In these cases, we can expect considerable variation between their cognitive roles, and so it is not at all clear whether Greenberg and Harman's account of content similarity is able to guarantee the co-referentiality of these BoKs.

For we should note that even though these three putative distinct BoKs are said to be used in similar cognitive processes (e.g., categorization, induction, deduction, analogy-making, planning, or linguistic comprehension, etc.), they are also said to be used in *very different ways*. For example, as Machery (2010) points out, while prototypes are assumed to be used in processes that compute the similarity between a prototype and other representations in a linear manner, exemplars are assumed to be used in processes that compute the similarity between a set of exemplars and other representations in a nonlinear manner. In turn, theories (e.g., a BoK involving some causal knowledge) are assumed to be used in "cognitive processes that are similar to the algorithms involved in causal reasoning" (p. 201). Thus, if the content of a BoK derives from its use, then prototypes, exemplars and theories associated with the same category are not expected to bear the similar contents that are supposed to guarantee their co-referentiality. We conclude from this that Greenberg and Harman's version of CRS cannot be successfully adapted to Machery's view of BoKs.

Referentialism and BoKs. As opposed to Intensionalism, Referentialism assumes that the extension of a concept determines its content. A prominent version of this view states that the content of concepts is determined by virtue of them being in a certain causal-cum-nomological relation to their referents. Here, Fodor's informational semantics is the best known position (Fodor 1987, 1990). According to this theory, the content of a concept qua primitive (or unstructured) mental representation is determined by the concept's standing in an appropriate causal-cum-nomological relation to things in the world. Motivated by the notion of content as information put forward by Dretske (1981), the theory is based on the idea that a concept carries information about a given property if the concept is under the nomological control of that property. More specifically, the idea is that types of concepts are causally connected to the properties their tokens express by means of there being a natural law that prescribes this (actual or counterfactual) causal connection. Thus, contrary to the alternative described in the preceding paragraphs, the content of a concept like APPLE would not be determined by its relations to other concepts (e.g., FRUIT, TREE, etc) but by a causal-cum-nomological relation that connects (tokenings of) the concept APPLE with (tokenings of) the property of *being an apple*.

The mention of Fodor's informational semantics is pertinent here since Machery himself resorts to this view when pressed by his critics to explain how his BoKs refer and how they can be said to be co-referential (Machery 2010, p. 235). He suggests that the retrieval from long-term memory of a given BoK (e.g., a prototype of squirrel) could be nomologically linked to the presence of a given property (e.g., *being a squirrel*) in a person's environment. Co-referentiality would then be explained by the fact that different BoKs could hold nomological links to the same property, as it would be the case with a SQUIRREL-prototype, a SQUIRREL-exemplar and a SQUIRREL-theory, for example. In addition, given that those BoKs are supposed to share the same informational content, the problem of (in)stability of content would then not be a threat to Machery's heterogeneous BoKs. However, I think there are good reasons why a causal view of content can't be made to work for Machery's distinct BoKs. Let us elaborate why.

The main problem is that this causal view does not seem to guarantee the co-referentiality of BoKs associated with the same category. There are situations where

prototypes, exemplars and theories about the same category could be said to be nomologically linked to things in the world that belong to different categories. To use a common example, consider the case of BoKs associated with the category of grandmothers. Prototypical grandmothers are gray-haired, have wrinkled skin and wear glasses. But there are grandmothers that do not satisfy these properties (e.g., Sarah Palin) and thus they fail to trigger a GRANDMOTHER-prototype to be used in cognitive processes that mainly depend on similarity to a prototype (e.g., prototype-based categorization). Instead, one would expect them to be linked to a BoK corresponding to a GRANDMOTHER-theory, the structure of which is likely to include the element of information that grandmothers are mothers of a parent. Consequently, a subject's GRANDMOTHER-prototype and GRANDMOTHER-theory can fail to co-refer because these BoKs can fail to be nomologically linked to the same referents.

To make this point clearer, consider the reverse case: an old woman that looks just like a prototypical grandmother but has never been a mother (e.g., Betty White). People will probably identify her as a grandmother through the activation of their GRANDMOTHER-prototype, while their GRANDMOTHER-theory takes no part in cognitive processes that mainly depend on, say, causal, functional or explanatory knowledge about a category (e.g. theory-based categorization). So we have another situation where a prototype and a theory about the same category are not co-referential (cf. Margolis and Laurence 2010).[52]

Suppose Machery might address these worries by adding that BoKs about a given category share a common informational core that guarantees their coreferentiality. Thus, it is possible to claim that such core is a common constituent of different BoKs about a given category and that it is this core that is nomologically linked to a property in the environment. In this sense, someone could think that the nomologically-based retrieval of any given BoK about *x* (e.g., GRANDMOTHER-

[52] Here a Fodorian might reply that this critique does not affect the referentialist view of the content of BoKs since, according to Informational Semantics, content is determined by the information carried by a concept, not by its inferential role (whether prototypical or theoretical). But in the context of Machery's view of BoKs, the information that is constitutive of a BoK is that which is activated by default in order to be used in cognitive processes, rather than by the role of that BoK in cognitive inferences (see Machery 2015). So, if content is identified with information as per Informational Semantics, BoKs whose respective default bodies of information fail to be co-activated by a given referent cannot be said to be under the nomological control of that referent and, therefore, they cannot be co-referential.

prototype) does not need to imply the retrieval of a whole different BoK about *x* (e.g., GRANDMOTHER-theory), provided that the shared core about *x* is retrieved every time one of those BoK is used in cognitive processes where they play a central role. However, this is a solution that does not work because it is incompatible with Machery's requirement that each distinct BoK about *x* corresponds to some default knowledge that is retrieved independently of that of other BoK about *x*. Recall Machery's distinction between context-independent default knowledge and context-dependent background knowledge mentioned in section 2. Even though distinct BoKs about *x* (e.g., CAT-prototype, CAT-exemplar and CAT-theory) are required to always overlap in different contexts, an individual is expected to activate *entirely different* default BoKs about *x*, not just some overlapping background knowledge about *x* that may be activated under certain contexts (e.g., that her cat is called *Lia*). In sum, to advance the existence of common informational cores between BoKs designating the same category cannot help to explain their coreferentiality, since the information required for those cores to be literally shared is likely to correspond to some background knowledge that is, by definition, not constitutive part of a BoK.

One final worry regarding the application of a referentialist approach to Machery's view of distinct BoKs has to do with whether these BoKs are construed at the right level of abstraction. As Schneider (2010) notes, if the semantic properties of co-referring BoKs can be individuated informationally, then the fact that prototypes, exemplars and theories conform to distinct processing properties does not challenge the idea that these BoKs could fall under the same concept type. Concepts could thus be understood as a superordinate—functional—kind (cf. Edwards 2010; Samuels and Ferreira 2010; Rey 2009; Lalumera 2013). Consider, for example, the view that the class of concepts is a natural kind that contains subordinate kinds including, e.g. prototypes, exemplars and theories in the same way that the class of minerals is a natural kind with subordinate kinds such as pyrite, topaz and quartz. In this sense, the causal view of content Machery thinks might work for his BoK really fixes the content of a superordinate kind, which, in turn, explains how any three different bodies of knowledge about a given category are actually coreferential.

At this point Machery might contend the notion of concept as a superordinate kind does no useful work in psychology and that, as he has argued, keeping this notion would impede scientific progress (Machery 2010, 2015). But this claim is

highly controversial, and many have pointed out that the notion of concept does have many explanatory virtues. Besides providing psychology with a notion to make interesting generalisations about high-level cognitive processes (Scarantino 2010; Weiskopf 2010), concepts have shown to be a useful heuristic in guiding psychological research (Hayes and Kearney 2010; Margolis and Laurence 2010). To elucidate this view, consider Cowie's account of "messy" or imprecise concepts (Cowie 2009).[53] In a related discussion about good and bad concepts, Cowie defends what she calls *vulgar pragmatism*, according to which imperfect and imprecise concepts should be preferred in the absence of better options when those concepts are part of a flourishing research programme.[54] She specifically raises this defence in connection with the concept of 'innateness', where vulgar pragmatism is aimed at answering the question of when and how the use of a particular messy or imprecise term can be justified. Together with the objections above, Cowie's idea of vulgar pragmatism can thus provide additional reasons to reject Machery's claim that *concept* hinders scientific progress.

7. Conclusions

Machery has argued that psychological explanation should renounce the notion of concept and focus on how different bodies of knowledge are used in the cognitive processes underwriting our higher cognitive competences. But an essential aspect of psychological explanation—content—is absent from Machery's alternative to concepts. In this paper, I examined possible options for the problem of how Machery's bodies of knowledge could attain their contents.

These options were based on the distinction between two broad approaches to content determination: Intensionalism, according to which the content of a concept determines its extension and Referentialism, according to which the extension of a concept determines its content. Idealised versions of known theories of content determination that we considered representative of the two main approaches failed to supply Machery's alternative to concepts with a viable account of content. In the case

[53] Cowie makes no distinctions of usage between 'concepts' and 'terms'.
[54] The scientific relevance of a certain kind of imprecise or vague concepts will be further developed in Chapter 6.

of Intensionalism, an understanding of BoKs in terms of the conditions to possess certain modes of presentation of a given referent proved incompatible with Machery's eliminativist claim regarding a general notion of concept. Alternatively, the problems to specify some minimum form of content stability for different BoKs in terms of conceptual roles stem directly from the models of conceptual representations they are based on. In the case of Referentialism, Machery's suggestion that Informational Semantics could be adapted to his proposed BoKs proved to be inadequate because this referentialist view does not support the conceptual heterogeneity Machery puts forward. My conclusion is that the problems to supply Machery's proposal with a viable notion of content constitute compelling reasons to resist his argument for concept eliminativism.

Chapter 5

Natural Kindness Misconstrued

1. Introduction

Chapter 3 illustrated a case where a concept eliminativist argues that the term 'concept' should be eliminated from scientific vocabulary on the grounds that what psychologists call concepts do not really form a *natural kind*. In this chapter I will defend the theoretical relevance of 'concept' by arguing for an appropriate understanding of natural kindness in the context of an immature science of the mind. Philosophers and natural kind theorists have proposed different characterisations of the notion of natural kinds which they find suitable either to support or reject their viability in a given field of enquiry. In my view, none of those conceptions is suitable to argue for or against the theoretical relevance of the notion of a concept in the scientific study of the mind.

In alignment with a dominant naturalistic tradition within contemporary philosophy of science, I take the term 'natural kind' to refer to the groups of particulars, properties (states, relations, etc.) in the world that the classificatory schemes of natural sciences are supposed to be about. One important assumption here is that these schemes facilitate scientific generalizations and are assumed to map onto the structure of reality in terms of worse or better approximations to it. The view that a mature science should successfully reflect real and theory-independent divisions in nature is called Scientific Realism and, in this chapter, I take this view for granted. This view is compatible with the idea that, in the case of the classificatory schemes of immature yet fruitful sciences, the methodological role of certain imprecise theoretical terms is to be preferred.

The chapter is divided into three main parts. The aim of the first part (section 2) is to briefly enumerate the most prominent characterisations of the notion of a natural kind and present the case of the disunity of special sciences. In the second part (sections 3), I will characterise the sources of an apparent tension underlying two opposing philosophical perspectives in the discussion of natural kinds. I will call them

AUSTERE and INDULGENT, respectively. In discussing the explanatory adequacy of natural kind terms, INDULGENT advocates tend to reject the accounts of natural kinds offered by AUSTERE advocates, and vice versa. With a focus on the study of concepts, the third part of the chapter (sections 4) will argue for the need of a mediating methodological solution and introduce some of the requirements for sketching it. Later on, in Chapter 6, I will present a possible solution that could help to overcome the tension between these two perspectives.

2. Natural kinds and the disunity of special sciences

2.1. Natural kinds: an elusive notion

Terms such as ‘carbon’ and ‘water’ are usually thought to refer to different types of stuff in the world. Likewise, terms such as ‘thistle’ and ‘tiger’ are usually thought to refer to different types of individuals in the world. A common assumption regarding the usage of those terms and the like is that their putative referents belong to different groups of objects with some putative interesting theoretical property (or set of properties) in common. Whether or not groups of objects like those are to be scientifically deemed as natural kinds is a topic of ongoing debate. However, what is clear is that not anything can count as a good candidate for a natural kind (e.g., compare the classes formed by different chemical elements with the class of things that taste bitter), and this is an important initial remark because it is objects that are likely to correspond to natural kinds that are scientifically interesting.

Scientific theories are expected to provide us with reliable knowledge about the structure of reality and they do so by postulating, among other things, theoretical terms that designate the existence of probable natural kinds relevant for discovering new generalizations and, thus, for gaining better understanding of the way things really work in nature. In this respect, to the extent that scientific theories are successful in achieving that goal, we are then justified in adopting a realist attitude towards the types of objects that those natural kind terms are said to refer to.

Different prominent reviews regarding natural kinds (e.g., Hacking 1991; Boyd 1991; Samuels 2009; Magnus 2012) suggest that, even though there is no

agreement about a clear set of conditions that a given category of objects has to satisfy in order for it to count as a natural kind, there are, at least, three recurrent conditions:

1. They are specified in terms of a set of properties that tend to co-vary
2. The properties specifying the natural kind are indicative of other scientifically interesting properties (i.e., they are properties that support inductive generalisations)
3. The properties specifying the kind are natural (e.g., they are to be discovered, not defined by people)

Philosophical accounts of natural kinds that appeal to conditions such as 1, 2 and 3 are typically aimed at providing a characterisation of what it is for something to objectively be a natural kind and, thus, a way to distinguish natural kinds from arbitrary or conventional categories (Bird and Tobin 2015; Brigandt 2011). In this sense, for example, chemical elements are better candidates for being natural kinds (likely to reflect some type of natural divisions in nature) than the class of things that taste bitter, because chemical elements (e.g., Cu), not bitter-tasting things, satisfy conditions 1 (e.g., the atomic structure of Cu is common to all instances of Cu), 2 (e.g., its atomic structure determines the chemical reactions Cu can participate) and 3 (e.g., the atomic structure of Cu is a matter of discovery, not a matter of convention). The putative natural kind status of groups of things that do not satisfy conditions 1, 2 and 3, such as the class of bitter-tasting things, can be thus put into question.

Notwithstanding the apparent advantages of this traditional approach to distinguishing natural kinds, there are many cases where it seems somehow reasonable to take something to count as a putative natural kind even if it does not satisfy some or all of the conditions above. Take, for example, the case of molecules which are artificially generated (e.g., ascorbic acid, dubnium); the case of biological species and higher taxa (e.g., the vertebrate/invertebrate distinction plays a prominent classificatory role in biology but those higher taxa seem to be less clearly a natural kind than particular species); the case of planets and asteroids where category membership seems to be a matter of convention (e.g., Pluto isn't any longer a member of what astronomers take to be the natural kind formed by planets), etc.

Philosophers have advanced different understandings for distinguishing objective natural kinds and their relevance in scientific theories. One proposal is based on the idea that natural kinds are the sort of categories involved in the true laws of nature (Magnus 2012). Consider, for example, the case where a true natural law is said to express a relation between universals (i.e., properties that different things can instantiate, as when Fido and Laika instantiate the universal property of *being a dog*). In this case, if, say, "all birds lay eggs" expressed a true natural law, then birds would designate a natural kind because that type of animals would correspond to a universal that participates in natural laws. Of course, what a true natural law really amounts to and whether or not available scientific generalisations might correspond to true laws of nature are open questions. But, for our present concern, it is interesting to note that this particular account of natural kinds seem to be consistent with at least two of the conditions stated above, namely condition 2 and 3, since this account allows for properties that support inductive inference (e.g., "lay egg" grounds inductions about all the members of the kind oviparous animals) and those properties are the result of discovery, respectively.

Another proposal to characterise natural kinds is essentialism, the view that objective natural kinds have essences. Essentialists have provided different conceptions of essentialism,[55] but it will suffice to illustrate this view in the light of one typical conception, namely one where it is assumed that for each kind K, there is an essential property P without which K cannot be the kind that it is. Even though the issue of which features on the world count as an essential property of any given kind is a matter of debate, essentialism allows for the essences of a kind to be qualified in different ways (e.g., structurally, relationally, intrinsically, etc.). Take, for instance, the cases of chemical kinds such as copper (Cu) and biological kinds such as Homo sapiens. Presumably, an essential property of Cu is its atomic structure, which is shared by all the members of the kind and explains in which chemical reactions instances of Cu can take part. Likewise, it is possible to characterise essentialism in relational terms (e.g., Mayr 1969), such as when, say, Homo sapiens are said to form a putative natural kind depending on whether they belong to the same reproductively isolated population. In this latter case, what is essential to the kind is its specific membership status. Another possibility is to argue, as Boyd (1991) has done, for a

[55] See Bird and Tobin 2015 for a more detailed discussion.

notion of essence in terms of a non-accidental correlation of properties. For instance, in alignment with this proposal, the species *Homo sapiens* could form a natural kind if it was the case that the members of this kind tend to share a set of causally grounded properties, even if there is no particular property within this set that is shared by all of the members (see next section, below).

A focus on relational properties may also be understood as an alternative to essentialism, such as in the case of functional kinds. Some philosophers (e.g., Quine 1969; Hacking 2007; Brigandt 2009) have pointed out the role of functional kinds in inductive generalisations, generalisations which may describe certain relations between the members of the kind, rather than certain intrinsic structural essence. For instance, as Brigandt (2011) has illustrated it, certain ecological generalisations can describe the categories of predator and prey in terms of the changes in the sizes of certain predator and prey populations. Likewise, certain biological generalisations can describe the category of vertebrates in terms of their philogenetic relation of descent from common ancestors' (e.g., the understanding of species defended by Cracraft 1983). Another example is money, understood as a multiply realizable functional category which plays an important role in many microeconomic generalisations (e.g., Gresham's Law and Thiers' Law). Regarding the conditions for natural kindness stated above, functional kinds specified in this way seem to satisfy condition 1 (in the form of a set of functional relations that tend to co-vary irrespective of how they are physically realized) and 2 (in the form of induction-supporting functional categories). But it is controversial to say that these kinds (as described above) can satisfy condition 3, since they do not provide a criterion for distinguishing categories whose functional properties are conventional from those which are not conventional. In this respect, however, some functionalist could respond that, even though 'money' designates a conventional category, the functional roles attributed to the members of the category are not really a matter of convention.

Another possible way of understanding natural kinds is in terms of the idea that natural kinds can be equated with property clusters. Particularly important for the present chapter is the proposal defended by Boyd (1991), according to which natural kinds are Homeostatic Property Clusters (HPC). As we will see, one of the advantages of this proposal is that it provides a solution that intends to reconcile the idea that

natural kinds have essences and the apparently incompatible functional characterisation of natural kinds described above.

2.2. Boyd's solution to Natural Kindness

One way to characterise natural kinds is in terms of some natural property that is shared by all the members of a given kind (e.g., Quine 1969). Another possibility, which is an expansion of the previous characterisation, is in terms of families of properties that are contingently clustered in nature (Bird and Tobin 2015). In this case, a relevant issue has to do with what it is that clusters such families of properties together. For instance, someone might have reasons to believe, as Hacking (1991) does, that what unites the properties of a given property-cluster kind is not a fact about nature, but human interest. Hacking's claim is based on the idea that natural kinds should be specified by some mind-independent set of conditions and the realisation that property-cluster kinds do not satisfy this condition. Alternatively, one might think, as Boyd (1991) does, that what unites property-cluster kinds is not human interest, but some kind of causal mechanism. In Boyd's words,

> On that conception a natural kind is associated causally with a large family of methodologically important properties. Even if the kind is thought of as being defined by a set of necessary and sufficient conditions its naturalness is a reflection of a wider sort of property correlation. It is natural to inquire whether in defining some kinds we might defer more fully to nature and take the kind in question to be defined by the larger family of correlated properties rather than by any special sub-set singled out as providing the necessary and sufficient conditions. If this possibility is acknowledged then it is reasonable to inquire whether there may be kinds so defined except that the relevant property correlations are not perfect, so that the set of correlated properties functions as a property- cluster. (p.141)

What Boyd (e.g., 1991, 1999) specifically claims is that natural kinds are *Homeostatic Property Clusters* (HPC). According to this conception, the members of a given natural kind are grouped together in virtue of a cluster of properties the co-occurrence of which is determined by some underlying causal mechanism. Importantly, none of the correlated properties in the causally-regulated cluster that the members of an HPC natural kind tend to possess has to be shared by all the members. In other words, whereas a certain cluster of properties can be said to be sufficient for

kind membership, none of the properties in that cluster can be said to be necessary for kind membership. This provides room for variation of the properties in the cluster, such as, for instance, when the cluster incorporates or ceases to include properties as a result of adaptive changes driven by environmental pressure (Bird and Tobin 2015).

Both the properties that may be part of a particular HPC natural kind and the causal mechanisms determining the homeostasis of the property cluster are an *a posteriori* theoretical question, in the sense that they are supposed to be the result of scientific discovery. In this respect, according to Boyd (1991), a paradigmatic case of HPC kinds are biological species, given their paradigmatic role in supporting scientific generalisations. Boyd is committed to the realist methodological assumption that, because natural kinds are discovered *a posteriori*, putative natural kinds are expected to play a significant epistemic role in the formulation of successful inductive generalisations. The realism involved in this assumption is reflected in the idea that the way in which we are to *accommodate* our categories to the objective causal structure of the world is a function of the success of both induction and explanation (Boyd 1991). Hence, his proposal that, in the case of biological species, while the reliable correlation of the properties associated with an HPC kind is what grounds instances of inductive inferences and explanations, the natural status of the natural kind depends on the extent to which its members are scientifically interesting (Brigandt 2011).

There are at least two types of homeostatic mechanisms relevant for defining biological species as HPC natural kinds: intrinsic and extrinsic (Bird and Tobin 2015).[56] An example of intrinsic homeostatic mechanism may involve gene exchanges within a population. Thus, for instance, people who are extremely tall (e.g., 2.5m) are rare cases because, given the intrinsically regulated property cluster defining the species, individuals with those genetically determined characteristics are less likely to reproduce than others. In turn, an example of extrinsic homeostatic mechanism may involve common selective factors in the niche of a species. Sometimes, organisms modify their ecological niches in ways that can alter the very relationship between the organisms and their relative niches, a process that can be said to have effects in the

[56] Another possibility, which I will not focus on in this chapter, could involve some continuous interplay between intrinsic and extrinsic homeostatic mechanisms, as it might be relevant to consider in light of recent proposal regarding the role of niche construction in evolutionary processes (e.g., Laland and Sterelny 2006).

evolutionary patterns of those organisms. For instance, beaver dam-building does not only favour the propagation of dam-building alleles (Dawkins 1982), but it is also said to influence subsequent beaver evolution by means of the selective pressure acting on many other beaver traits as a result of the inherited modified selective environment (e.g., Laland and Sterelny 2006). If this *niche-construction* evolutionary perspective is correct, then deviations will be expected to be selected against in environments that can be modified by niche construction. In both cases, the homeostatic mechanisms can be said to be an instance of self-regulatory process in the sense that these mechanisms secure the presence of many properties in the cluster, along with minimising deviations from the cluster. As Hacking (2007) puts it:

> The species is in equilibrium in the sense that descendants that diverge too far from the cluster of properties die out or form a new group. Species thus endure thanks to a network of causes that produce stability of a homeostatic sort. (p. 235)

Note that HPC kinds can be said to satisfy all of the conditions for natural kindness stated in the previous subsection. Condition 1 is satisfied because the property cluster associated with the members of a putative natural kind is secured by its sustaining causal mechanisms. Condition 2 is satisfied because the presence of a given set of properties in a HPC kind is constrained by the extent to which those properties satisfy induction and explanation. In other words, for a property cluster to count as a putative natural kind it has to be involved in induction and explanation. Finally, condition 3 is also satisfied, since HPC kinds are ultimately specified, not by convention, but *a posteriori* so as to reflect the actual causal structure of the world.

Note that, as I have mentioned in Chapters 3 and 4, this is the notion that the eliminativism regarding concepts put forward by Machery (2009, 2010) employs. This eliminativist proposal is a case of eliminativism where a given putative natural class (e.g., the class of concepts) is rejected on the grounds that the theoretical term (e.g., 'concept') we used to designate it is too vague or confused to be scientifically interesting. Machery's claim is that the notion of concepts does not really pick out a natural kind because this notion misleadingly conflates three distinct bodies of knowledge (viz., prototypes, exemplars and theories) which, as opposed to what he

thinks about concepts, are said to be independently sustained by causally active internal mechanisms. I will come back to Machery's eliminativist view in later sections since rejecting this type of concept eliminativism by defending the theoretical status of concepts is central to the present work.

2.3. The (dis)unity of special sciences

At the beginning of the chapter, I briefly introduced scientific realism as the view that the classificatory schemes of natural science is ultimately aimed at mapping onto the objective structure of reality. One thing this view presupposes is that the way the world is really structured is independent of the way scientists might think it is structured, in the sense that the reality of the world (i.e., the way things actually are in the world) is independent of the theories that scientists develop in order to understand it. On the face of it, however, different special sciences (e.g., biology, psychology, geology, economics, etc.) seem to be about kinds (e.g., species, mental states, tectonic plates, services, etc.) which can't be clearly said to be theory-independent in the same way. In some cases, the desideratum of theory independence is highly controversial, even if the theorised kinds play a role both in induction and explanation (e.g., social mobility, stereotype threat, monetary exchanges, etc.). The question arises about the extent to which the kinds that special sciences theorise about satisfy proposed conditions for natural kindness and, if they do not satisfy any known conditions (thus compromising the established commitment to scientific realism), what it is one could say about their theoretical credentials.

In what follows, I will focus on two main issues. First, I will briefly survey different examples where some of the kinds that certain special sciences take for granted seem to be at odds with scientific realism (in the way I have broadly characterised it). The second issue to address is a version of the thesis of the unity of science, according to which the theorised natural kinds of all true theories of special sciences should ultimately reduce to those of some basic science. Let us take a look at each of these two issues in turn.

Chemical kinds are a good starting point because these kinds are normally considered to be paradigmatic instances of natural kinds (LaPorte 2004; Magnus

2012). When I introduced essentialism in 2.1 above, I mentioned the case of copper (Cu) as a putative case where it is essential to something to be constituted by a given property. In the case of Cu, a good candidate for such essential property is its atomic structure, in the sense that whatever object which is only made up of atoms with 29 protons in their nuclei can't be said to be made of anything else except copper. However, this microstructuralist view does not seem to be useful when it comes to distinguishing between chemical compounds for, at least, two reasons. First, chemical compounds are normally identified by their constituent elements (not their microstructure) such that H2O can be said to be what it is because the compound is constituted by two hydrogen atoms and a single atom of oxygen. Second, there are compounds elements which are distinct compounds even if they share the same constituent elements in the same proportions. For example, fulminic acid and cyanic acid have the same constituent elements (HCNO) but they different in certain physical properties, namely the spatial arrangement of the atoms their constituent atoms (Bird and Tobin 2015). Indeed, many scientifically interesting properties of certain substances (e.g., melting and boiling points of water) can't be explained merely according to microstructural criteria (Weisberg et al. 2011).

Thus, it should come as no surprise that the problems of *microstructuralism* (i.e., the view that only microstructural properties can individuate chemical kinds) in the philosophy of chemistry have motivated extensive discussions about the natural kind status of many chemical kinds. For instance, Needham (2000) proposes a view of macroscopic objects (e.g., a body of water), and, given the empirical relevance of some of the dynamical aspects of certain chemical compounds, he concludes that:

> A macroscopically oriented account of sameness of kind doesn't challenge the claim that quantities of water have some appropriate range of microfeatures under specified conditions. But recognising microproperties is not to favour them as more essential than others. (p. 21)

From this quote, the suggestion is that the microstructure of a macroscopic objects *O* is not what makes *O* to be what it is. Alternatively, Hendry (2006) agrees that objects such as a body of water should be viewed as macroscopic objects which are both dynamic and heterogeneous at the molecular level (owing to his view that

water has H2O as its ingredients), but rejects the idea that *being water* should be individuated in terms other than as some arrangement of H2O molecules.

Let us turn to the philosophy of biology. Discussions of natural kinds in the philosophy of chemistry should make it easier for us to understand why it has been so difficult for philosophers to provide an account of biological species and how it is that different species can be distinguished. Species evolve and this fact raises the problem of when to distinguish new species form their ancestors. Moreover, because species are normally classified in relational terms (i.e., in terms of their common ancestral descent), notions of natural kindness that seem useful in other more basic sciences (e.g., the essentialist notion of natural kinds typically used in Chemistry) do not seem to be relevant for characterising species as natural kinds. Different species concepts are on offer (e.g., the concept of species as isolated interbreeding populations offered by Mayr 1969 and the phylogenetic concept of species put forward by Cracraft 1983), but the norm among philosophers interested in the problem of species is to provide examples that illustrate intuitive conceptions, rather than a clear analysis of what they mean by their putative natural kind concept of 'species' (LaPorte 2004). Thus, while 'tiger' and 'mammal' can be said to be frequent examples used to illustrate those intuitive conceptions of species, they are also examples used to illustrate disagreement about what could actually reflect real divisions in nature. Indeed, philosophers usually make a distinction between taxa (e.g., the class of tigers) as reflecting real divisions and higher taxa (e.g., the class of mammals) as reflecting conventional divisions (e.g., LaPorte 2004; Bird and Tobin 2015).

Debates regarding natural kinds in biology have also focused on whether species are kinds or individuals. To say that certain organisms should be viewed as individuals is the same thing as saying that those organisms are *parts* of species, rather than *members* of a species as a natural kind (LaPorte 2004; Brigandt 2009). One common motivation for the view that species are individuals involves certain unwanted consequences of the evolving character of organisms. For example, if organisms change their characteristic properties over time then the requirement that members of natural kind tend to share common intrinsic properties is no longer relevant for characterising species as a natural kind (Bird and Tobin 2015). The view of species as individuals is intended to object to the inference whereby realism of species is rejected from the (purported) conclusion that species are not natural kinds.

However, this is not the only way to reject the view that species do not reflect real divisions in nature. For example, LaPorte (2004) has argued that both conceptions (i.e. species as kinds and species as individuals) are not incompatible with one another on the grounds that certain further refinement of both interpretations "will not affect the acceptability of ordinary scientific claims" (p. 17). In any case, both this compatibilist view and the view that species are individuals require abandoning the essentialist idea that species must have certain intrinsic properties in common (e.g., Putnam 1975), which in turn, is said to be relevant for species to take part in laws of nature (Bird and Tobin 2015). Two ways in which philosophers have tried to do without this traditional idea of essentialism are, first, by postulating some notion of essentialism which is not based on intrinsic properties (e.g., one that is historical or relational properties) and, second, by offering alternative pluralistic views natural kinds, where different conceptions are accepted on the basis of their theoretical merits (e.g., Dupré 1993).

What about psychology, the scientific study of the mind? The question of whether or not certain mental states form a natural kind is central to this chapter and I will attempt to sketch the basis for a response later on. For the time being, suffice it to highlight some of the problems of providing a positive response to this question in the field of Cognitive Science, where a basic intuition among theorists is that the mind (and therefore our mental states) is a causal factor of human intelligent behaviour. Ever since the emergence of Cognitive Science as a reaction to the then dominant radical behaviourism,[57] different understandings of mental kinds have been offered.

In an attempt to develop a materialist theory of mind that allows for mental causes, central-state identity theorists maintained that mental events (states and processes) are identical with brain events and that being in a certain mental state (e.g., believing that the train is late) is identical to being in a certain brain state (Smart 1959). This identity theory of the mind takes one of two forms, namely token physicalism (or *token* identity theory of the mind) and type physicalism (or *type* identity theory of the mind). Whereas the former typically maintains that tokens of mental state (e.g., my current desire that it snows) are identical with tokens of neurophysiological state, the latter typically maintains that types of mental state (e.g.,

[57] See Miller (2003) on the emergence of Cognitive Science.

having the desire that it snows) are identical with types of neurophysiological state. Another important way in which these two doctrines of physicalism differ is that type physicalism, but not token physicalism, rules out the possibility that anything which does not have neurons can have mental properties (Fodor 1981). In this sense, what type physicalism rules out is *multiple realizability*, i.e. thesis that a single mental state can be realised by physically diverse kinds. However, it is a logical and empirical possibility that distinct physical creatures (e.g. machines, silicon-based Martians, etc.) can be in the same kind of mental state.

Another substantial assumption guiding much of research within mainstream Cognitive Science is that our brains are information processing systems, an assumption that, contrary to type physicalism, is compatible with the possibility of multiple realization. If type physicalism theorists are wrong, then it could be argued that, because mental states are not clearly reducible to brain states, putative mental kinds are good candidates for eliminativism. For instance, it could be argued that scientists should eliminate the notion of (irreducible) mental kinds because this notion is motivated by a common sense theory of the mind which is a hopelessly mistaken theory (cf. Churchland 1981, 1988). On the contrary, another option is a theory of mental states which is materialist yet one where metal states can abstract from the physical structure of the creatures that bear them. Functionalism is the view the mental states are individuated by their functional role, irrespective on which physical kind they are realized. In this sense, functional mental states are irreducible, yet compatible with token physicalism since it is possible that neurophysiological events are the only thing with the appropriate functional properties that individuate mental kinds (Fodor 1981). It is also compatible with the physicalist demand that mental causation is also physical causation, yet, as Bird and Tobin (2015) points out, it appears to violate the traditional requirement that members of a natural kind (notably, their physical realisations) should share some natural properties in common.

In the case of social sciences, the issue of what actually count as theory-independent natural kinds is even more controversial that in the other cases discussed so far. It is not clear whether there are any social kinds at all and, if there were any (say, e.g., marriage, gender, unemployment and the like), it is also unclear to what extent they could be compatible with the naturalist and realist doctrines that underlie natural science. One of the reasons why social kinds are said to be incompatible with

the idea of natural kinds is that social kinds are inevitably dependent on human interest and attitudes and, thus, they are some species of social construction (e.g., Hacking 1999). However, similar worries have been raised with respect to many traditional conceptions of natural kindness. For instance, some (e.g., Khalidi 1998) have argued that social kinds and natural kinds only differ from one another in the sense that social kinds are part of classificatory schemes that are more dependent on human interest than in the case of the classificatory schemes of so-called natural kinds.

It is also possible to claim that social kinds are some kind of HPC because they tend to share a given properties cluster *PC*, but then the question arises as to how the co-variation of the properties in this cluster is sustained. It is inherent to the formulation of the notion of a HPC that the properties in the *PC* do not co-vary accidentally, but causally. So, perhaps, social kinds fail to be HPC because there is no (causal) mechanism that secures its homeostasis.

One way in which one can argue for the legitimate theoretical status of the putative natural kinds relevant to the taxonomy of many special sciences is by challenging the reductivist thesis that true or successful theories in special sciences should eventually be reduced to physics or some similar basic science. Fodor (1974) has notoriously argued against this thesis (at least, in its strongest version) and defended what he calls the disunity of science as a working hypothesis. The way he characterises natural kinds is in terms of the kinds that figure in natural laws. Hence, the way he characterizes the sort of reductivism he rejects is in terms of the necessary and sufficient condition that all the laws of a given special science be reducible to the laws of physics (or some other basic science, as it might be). In doing so, Fodor (1974) rejects a strong version of the generality of physics:

> […] reductivism entails the generality of physics in at least the sense that any event which falls within the universe of discourse of a special science will also fall within the universe of discourse of physics. (p. 101)

Thus, assuming that a given science corresponds to the formulation of a collection of laws,[58] he specifically objects to the assumption underlying reductionism according to which special sciences are to be reduced to more basic sciences by means of bridge laws connecting kind predicates from the reduced sciences with kind predicates of the reducing sciences. Bridge laws can be of two types, i.e., they can either express property identities or event identities, and Fodor thinks neither of them has tolerable consequences—these consequences being that every natural kind *is* a physical natural kind or that every natural kind *is coextensive* with a physical natural kind.

The point is that there are clear cases in special sciences which, if taken seriously, challenge the prospects of reductivism required by unity of science hypothesis to which the strong version of the generality of physics is committed. It is unlikely, for instance, that physical predicates can subsume the events that presumably fall under the laws of monetary exchanges in economics (e.g., 'x is a monetary exchange' iff x is physical natural kind subsumed by a physical description), since these types of events involve a number of highly disjunctive physical descriptions (e.g. wampum, cattle, cowry shells, banknotes, strings of digital code, etc.). Thus, the generalisations that are interesting in economics and physics involve events with physical descriptions that have nothing non-accidental in common, which, as Fodor (1974) argues, supports the thesis that economics is not reducible to physics:

> The point is that monetary exchanges have interesting things in common; Gresham's law, if true, says what one of these interesting things is. But what is interesting about monetary exchanges is surely not their commonalities under *physical* description. A natural kind like a monetary exchange *could* turn out to be co-extensive with a physical natural kind; but if it did, that would be an accident on a cosmic scale. (pp. 103-104)

Fodor claims that it is implausible that there are putative natural kinds in economics with co-extensive putative natural kinds in physics and this claim—he argues—is not special about economics but commonplace in the case of all other

[58] More precisely, a collection of theoretical predicates the satisfaction of which allows for certain observed events to fall under the laws of that science.

special sciences. Indeed, his main concern is to show that the thesis that neurological natural kinds are co-extensive with psychological natural kinds (i.e., the strong doctrine known as type-to-type physicalism) is wrong, on the grounds that, even though every psychological event can be paired with some neurological event, psychological events of the same kind could very well be paired with distinct kinds of neurological or non-neurological events (i.e., the weaker doctrine known as token physicalism).

This latter weaker doctrine allows for a reinterpretation of scientific reduction to which I feel inclined to adhere. According to this alternative understanding, the goal of reduction is an account of the (*heterogeneous* and *unsystematic*) physical mechanisms by means of which events conform to the laws (notably, counter-factual supporting generalisations) of the special sciences—not to find natural kind predicates in reduced and reducing sciences which are co-extensive with one another. In this sense, for example, even though natural kind predicates in chemistry (e.g., Cu, HNCO) are not to be seen as reducible to co-extensive natural kinds predicates in physics (e.g., subatomic particles), putative natural kinds in chemical events that fall under law-like generalisations can very well be compatible with the generality of physics.

I endorse the case for the disunity of special sciences (as a working hypothesis) and Fodor's suggested liberalisation of the generality of physics as a constraint for the acceptability of special sciences. The assumption that psychology as an immature scientific field is an irreducible special science will be instrumental to the qualified application of Boyd's notion of HPC kinds I intend to set forth in the next chapter. However, more needs to be said about the need of an alternative characterisation of the notion of natural kinds in psychology. Thus, in the remaining part of this chapter, I want to characterise the sources of an apparent tension between two dissenting philosophical perspectives on the notion of a natural kind resulting from discussions about an adequate construal of this notion. After that, since my main concern is the domain of the mental, I will argue for the need to overcome such tension on the basis of a general characterisation of the notion of a natural kind in the study of concepts that addresses the kind of worries that motivate the disagreements between the two perspectives.

3. Two approximations to the problem of natural kinds

If the construal of the notion of a natural kind to which one may be committed is too demanding (e.g., by appealing to some form of strict intrinsic essentialism), then it could apply (if at all) to much fewer cases than we may want and, in turn, it could lead one to overlook interesting generalisations in scientific fields where such construal is not relevant. On the contrary, if our construal of the notion in question is too permissive (e.g., by giving up the generality of a basic science like physics or by abandoning its role in law-like generalisations), then it could apply to so many cases that it would be difficult to tell the extent to which scientific taxonomies are informative regarding the general and objective character of reality. In this section, I will provide an idealised characterisation of two general perspectives on the problem of developing an adequate account of natural kinds, each of which aims towards one of the two mentioned situations. Call these perspectives AUSTERE and INDULGENT, respectively. In some respect or another, different accounts of natural kinds can be said to be inclined either towards AUSTERE or INDULGENT.

One useful way to proceed is by characterising how AUSTERE and INDULGENT comport with respect to a motley collection of familiar background assumptions that constrain the theorising about natural kinds (either explicitly or not). Different degrees of commitment to these assumptions can help distinguish theorists' tendency to adopt a perspective either towards AUSTERE or INDULGENT. It will be helpful to think of each of these two perspectives as adopting the extreme positions of a continuum between (what appears to be) a set conflicting assumptions. Depicting these two approximations in this way will hopefully make it manifest that there is a tension between them and that this tension is not resolved by simply choosing one of the two options.

(1a) *The objective/subjective assumption*. One of the reasons why philosophers have been interested in working out an account of the notion of a natural kind is the assumption that a correct account of this notion should help distinguish between (relatively) objective descriptions of the structure of the world from more subjective or conventional descriptions (Bird and Tobin 2015). A common agreement in an influential tradition about the notion of natural kinds is that not anything can

count as natural kind (whatever they turn out to be) because, if there are any natural kinds at all, it is a fact about nature that they are the natural kinds that they are (Hacking 1991). But that agreement is not shared in other traditions where theory independence is not deemed a requirement for some type of things to count as a natural kind. Thus, for example, dissenting positions that challenge the theory independence requirement for natural kindness might argue that natural kinds are metaphysically dependent on social conventions and our psychological capacities for recognising them. If the objective/subjective assumption is depicted as a continuum between more/less objective and theory independent accounts, AUSTERE can be said to aim towards the more objective end and INDULGENT towards the other.

(1b) In connection with the mental, a similar distinction can be depicted as a continuum between, on the one hand, the view that what is essential to our thoughts is their relation to the things in the world that they represent and, on the other hand, the view that what makes our thoughts special is their relations to our actions (i.e., to what we know how to do). A radical example of the first view is the Language of Thought Hypothesis (e.g., Fodor 2008). A radical example of the second is the view that thinking is always action-oriented such that what we think is *constituted* by what we can do (e.g., Prinz and Clark 2004). Call these two types of views Semantic Representationalism and Semantic Pragmatism, [59] respectively. Semantic Representationalism is consistent with the objectivist view that what makes our thoughts true or false is the way the world is. In contrast, because Semantic Pragmatism is often used as an umbrella term for a variety of views which are grouped together in a rather loose way (e.g., Prinz 2011), it is hard to clearly state a common objectivist commitment associated with all those views. But in any case, it is safe to say that, to the extent that Semantic Representationalism, as the term suggests, focuses on thought and a representational relation between its content and the world, Semantic Pragmatism comprises some type of reaction against Semantic Representationalism. In particular, the radical reaction I want to highlight is the claim that the concepts that allow us to have thoughts about the world are action-oriented, a pragmatist view about concepts that invites us to accept that what we know about the things that fall under the extension of a concept (e.g., knowing that trees can produce

[59] Semantic pragmatism is the term used by Fodor and Pylyshyn (2015) in their response to Prinz and Clark (2004).

shade) is constitutive of that concept (e.g., TREE). The reason for this, as the semantic pragmatist might argue, is that the content of a concept such as TREE (i.e., that which the concept TREE is about) is determined by the way in which we would act if we were confronted to things that are trees (e.g., look for sun safety on hot summer days). In this sense, whereas this strong Semantic Pragmatism is (explicitly or implicitly) committed to the subjectivist idea that the content of concept C is constituted by many of the things one might idiosyncratically know about C (notably, what we may happen to know about people's action-oriented behaviours with respect to, say, things in the extension of C), a strong Semantic Representationalism can be said to be committed to the objectivist idea that the content of C is completely unaffected by our beliefs about the things falling under C.[60] Thus, if the objective/subjective assumption is depicted as a continuum between some form of strong Semantic Representationalism and some form of strong Semantic Pragmatism, AUSTERE can be said to aim towards the representationalist end and INDULGENT towards the other.

(2a) *The ontological/methodological assumption.* Our understanding of the notion of a natural kind can also be characterised in terms of where one stands with respect to one of two interpretations of naturalism (roughly, the doctrine that our investigation into the nature and structure of reality should be free from "supernatural" elements). Traditional accounts of natural kinds are said to aim at providing a metaphysical characterisation of such notion (e.g., in terms of an answer to the question of "what is the nature of X?"). This interpretation contrasts with a recent focus on accounts that favour some kind of methodologically-oriented characterisations (e.g., Brigandt 2011; Pöyhönen 2013). In view of that, some people (e.g., Papineau 2015) broadly distinguish a commitment to the doctrine of naturalism along the lines of two opposing interpretations, namely, *ontological* and *methodological naturalism*. While ontological naturalism is said to appeal to philosophical argument and is preferentially committed to some physicalist view about the content of reality (thus providing an *a priori* constrain for natural kindness),

[60] To make both the objectivist character of Semantic Representationalism and the subjectivist character of Semantic Pragmatism more explicit, compare it to two different ways in which cognitive scientists tend use the term 'representation', namely, the *existential* and the *purely intentional* (Rey 2005). The former use expresses the objectivist notion of concepts whose putative content is in "the real world" (e.g., STONE) and the latter expresses the subjectivist (and, probably, circular) notion of concepts whose putative content is "in the head" (e.g., MEDUSA).

methodological naturalism is said to be preferentially committed to the authority of the scientific method in philosophical practice. An effect of this latter commitment is an emphasis on the scientific fruitfulness of natural kind terms. Given this epistemic focus, methodological naturalism—as opposed to ontological naturalism—is permissive towards highly dissimilar construals of the notion of a natural kind, provided that those construals are based on empirical considerations (e.g., Dupré 1993). If the ontological/methodological assumption is depicted as a continuum between ontological and methodological naturalism, AUSTERE can be said to aim towards the ontological end and INDULGENT towards the other.

(2b) We can also recognise this second assumption in the study of the mental. Take the case of the most prominent competing theories of concepts. These theories can be broadly classified in at least two groups, namely those that are mainly occupied with ontological issues about the nature of concepts (e.g., Fodor 1998; Peacocke 1992) and those where theorisation about concepts is heavily grounded on empirical evidence obtained by reliable procedures (e.g., Murphy 2002). Given the diversity of mental structures postulated by current psychological theories of concepts, some philosophers (e.g., Weiskopf 2009; Machery 2009) in the second group have even suggested that there is no single mental structure that can capture the causal and explanatory role that most concepts theorists in both groups expect concepts to play. Consequently, they defend the idea that a number of different mental structures (e.g., exemplars, prototypes and theories) can do that job. Still more radically, Machery (2009) has claimed that theories of concepts in psychology and philosophy are not converging theories because psychologists and philosophers do not really talk about the same thing when they discuss what they call concepts. Even if one does not accept such a sharp divide, it is safe to say that concept theorists can have very different interests, notably, while some of them may be primarily interested in, say, the problem of the nature of concepts (i.e., the answer to the question of "what are concepts?"), others are primarily interested in the explanatory role of concepts with respect to certain higher cognitive capacities (e.g., categorisation). If the ontological/methodological assumption depicts a continuum between a focus on the ontology of concepts and a focus on the empirically-grounded cognitive mechanisms supporting certain higher cognitive capacities, AUSTERE can be said to aim towards the ontological end and INDULGENT towards the other.

(3a) *The realist/anti-realist assumption*.[61] Within philosophy of science, a default position regarding the nature of our scientific knowledge is so-called scientific realism, i.e., a positive epistemic attitude towards the theoretical terms and descriptions of our best scientific theories. Realism regarding theoretical terms has been discussed by different authors under different characterisations (Chakravartty 2015). For current purposes, I will characterise a strong realist position about natural-kind terms as a commitment to the view that there is no significant difference between the referents of the natural-kind terms used in a current special science and the kinds in the taxonomy of a (future) completed version of that special science—cf. the notion of 'deep realism' in Magnus (2012). In this view, the former are just a good approximation to the latter. Call this position Strong NKT Realism. By contrast, anti-realist positions regarding natural-kind terms may hold a commitment to the view that natural kind terms need not (or, in its most radical version, should not) reflect objective natural kinds because their scientific merit is only of a practical character—as in the distinction Churchland (1985) makes between natural kinds and 'merely practical kinds'. In its strongest version, anti-realism may take the form of relativism or social constructivism regarding the meaning of theoretical terms. Call this latter position Strong NKT anti-realism. If the realist/anti-realist assumption depicts as a continuum between Strong NKT Realism and Strong NKT anti-realism, AUSTERE can be said to aim towards the realist end and INDULGENT towards the other.

(3b) Some version of this third assumption underlies important debates in the field of cognitive sciences. Consider the term 'intentionality' in the context of the debates about the naturalisation of the property that our minds have of being about or representing things, states of affair, etc. As it has been suggested, possible engineering solutions to this problem may involve specifying mechanisms, processes or relations which secure the mind-world relation (e.g., Dretske 1981; Fodor 1987). Call this type of proposals Aboutness Realism, as they are committed to *intentional realism*, the view that 'intentionality' designates some type of thing that is part of the natural order of things. Other options may involve solutions that do not endorse intentional realism. For instance, it is possible that the reason why it is so difficult to show how a physical

[61] Note that (3a) differs from (1a) in that (3a) involves an epistemological assumption about theoretical terms, whereas the (1a) involves an ontological assumption about the putative referents of those terms.

system can have intentional states is simply because there are no such states. Someone who thinks so might claim that there are no *intentional* states, without challenging the existence of mental states couched in some other way (e.g., Stich 1983). Others may go even further and take an instrumentalist stance that makes use of the descriptive apparatus of intentional explanation without committing to a realist position about it (e.g., Dennett 1987). Call this second type of proposals Aboutness Anti-realism, as, in this case, the term 'intentionality' is deemed to be empty. If the criterion of Intentional State-terms attitude is depicted as a continuum between Aboutness Realism and Aboutness Anti-realism, AUSTERE can be said to aim towards the realist end and INDULGENT towards the other.

(4a) *The unification assumption*. According to a strong reductive model of the unity of science, the putative natural kinds about which special sciences generalise should eventually reduce to those putative natural kinds that natural-kind terms in more basic sciences designate. One way in which this strong reductive model can be rejected is in terms of a weaker version of it where the reductionist requirement undergoes some liberalised re-interpretation. This is what Fodor (1974) does. Another way of rejecting that strong reductive model is by defending anti-reductivism and a conception of the disunity of science that is not committed to the generality of a basic science like physics. This is what Dupré (1993) does. Following Dupré, whereas the strong reductivist model of scientific unification is constrained by the metaphysical assumption that there is a unique and systematic order in nature, anti-reductivist models of the disunity of science need not accept such assumption. His proposal is a *pluralistic epistemology*, the idea that science should include a variety of projects of enquiry—which may or may not correspond to paradigmatic scientific disciples—merely guided by certain set of epistemic virtues (e.g., sensitivity to empirical facts, plausible background knowledge, coherence of knowledge, etc.). If the unification assumption depicts a continuum between a strong reductive model of the unity of science and a strong anti-reductivist models of the disunity of science, AUSTERE can be said to aim towards the reductivist end and INDULGENT towards the other.

(4b) Philosophers and cognitive scientists have extensively discussed issues connected with the autonomy of psychology and whether psychological theories and

phenomena are reducible to theories and phenomena in neuroscience. In general terms, there are at least three main positions with respect to reductionism of psychology. Some (e.g., Bickle 2003; Churchland 1981, 1989) have defended the view that there is no room for propositional attitude psychology in a mature science of cognition, where a mature science of cognition is to be understood in terms of a certain future developments in neuroscience. This is a view that clearly excludes the possibility of psychology as an autonomous science. In defence of the independence of psychology, others (e.g., Putnam 1967; Fodor 1974) have appealed to functionalist explanations of mental phenomena and the related notion of *multiple realizability*, i.e., the thesis that a given functional psychological state can be realised by indefinitely many structural kinds that perform the same function. It would appear that the autonomy of psychology is incompatible with any form of reductionism. However, philosophers have also looked for some kind of compatibilist solution as a third way. For example, Bechtel (2008) has recently defended the view that some variety of mechanistic explanations can be both *reductionist* and *emergentist*,[62] such that, while the decomposition of the component parts of a whole system is compatible with reductionism, the higher-level organisation and operation of the whole system secures an independent theoretical level. Putting aside any possible objections to this latter mechanistic solution, Bechtel's proposal illustrates the need to bridge two radically opposed positions regarding reductionism and scientific unification with respect to the mental. If the unification assumption depicts a continuum between some strong anti-autonomism regarding psychology and some strong autonomism regarding psychology, AUSTERE can be said to aim towards the anti-autonomist end and INDULGENT towards the other.

(5a) *Quantitative assumption regarding inductive generalisations*. A standard assumption about natural kinds is that one can make inductive inferences about them (Magnus 2012) and this fact raises the "how many" question, i.e., the question of how many generalisations might be sufficient for a given theoretical term to count as a legitimate natural-kind term. Note that the question does not directly depend on the *projectibility* of natural-kind predicates, which, in turn, depends on

[62] Bechtel (2008) characterises this notion as the behaviours that whole systems exhibit, which "goes beyond the behaviours of their parts." (p. 129).

whether the natural-kind terms of a science actually designate genuine natural kinds.[63] Instead, the question is about the justification of a *putative* natural-kind predicates, notably, the predicates expressed by the natural-kind terms of an immature science. If the inductive success of projectible predicates in a certain mature science is determined by the fact that those predicates are about genuine natural kinds, it appears reasonable to think that scientists are justified in inferring that natural-kind terms supporting many inductions are more likely to pick out genuine natural kinds than those supporting few inductions. However, the requirement of inductive generalisations might be less central (and, perhaps, even unwanted) in the case of immature sciences, where both informative theoretical terms and related taxonomies are expected to be highly imprecise. Thus, it is possible to distinguish scientific scenarios where the projectibility of natural-kind predicates is, say, a central requirement and where it is not. If the quantitative assumption regarding inductive generalisations depicts a continuum between the view that legitimate natural-terms are to support many inductions and the view that legitimate natural-kind terms do not have to support many inductions, AUSTERE can be said to aim towards the former and INDULGENT towards the other.

(5b) Some have defended the strong requirement the legitimate natural-kind terms in psychology should support a rich set of causally-grounded inductive generalisations and, on those grounds, they have rejected the scientific relevance of theoretical terms that do not seem to pick out kinds that support many scientific inductions (e.g., Machery 2009). The question of how rich or large the mentioned set of generalisations required for a legitimate natural-kind term must be can vary depending on the conception of natural kinds one is committed to. If it is part your preferred account of natural kinds that natural-kind terms must feature in natural laws (e.g., Fodor 1974), it might be difficult to provide an account of putative natural kinds supporting a rich set of inductions in psychology because, in general, there aren't any of those theoretical terms—At best, natural-kind terms in psychology can be said to feature in certain type of lawlike generalisations (e.g., *ceteris paribus* generalisations). By contrast, if your preferred account of natural kinds is so broad that, say, even artefacts can qualify as bona fide natural kinds (as in Machery's case), then the

[63] Natural-kind terms express projectible predicates if they enter into successful inductions, on the assumption that these successful inductions are ultimately confirmed by the fact that those terms designate genuine natural kinds.

chances are higher that a given putative natural kind can be said to support many scientifically interesting inductions. Thus, if the quantitative assumption regarding inductive generalisations depicts a continuum between the strong view that psychological kinds should support a rich set of inductive generalisations (perhaps, due to a permissive conceptions of natural kinds), on the one hand, and, on the other, the view that psychological kinds do not have to support many inductive generalisations (perhaps, due to a restrictive conceptions of natural kinds), AUSTERE can be said to aim towards the former and INDULGENT towards the other.

As I mentioned earlier, AUSTERE and INDULGENT are meant to characterise two idealised perspectives to the development of an adequate account of natural kinds. Surely, it would be too simplistic to say that each of the available theories of natural kinds strictly conforms to either one approximation or the other. Rather than that, I prefer to say that any given construal of the notion of a natural kind can be roughly associated to one or the two perspectives in one or several respects (see summary diagram 1, below). It can very well be the case that certain construal of natural kindness is committed to AUSTER in one respect and INDULGENT in a different respect. For instance, whereas Machery's view of natural kinds can be said to be INDULGENT with respect to the ontological/methodological assumption (e.g., consider the different types of empirically-grounded natural kinds he thinks can account for categorisation as a diverse family of similar processes), it can also be said to be AUSTERE with respect to the quantitative assumption regarding inductive generalisations (see, e.g., Hill 2010 and Machery 2010, respectively). The point to stress is that theorists choose to steer their accounts of a natural kind towards either AUSTERE or INDULGENT on the assumption that such choice is always done at a cost. Put differently, what seems to be at stake is whether the benefits of opting for one approximation or the other (in a given respect) outweigh the costs. However, as I want to maintain, that analysis is only pertinent in case the tensions I have characterised in terms of AUSTERE and INDULGENT (in some particular respect) are really incompatible with one another. Since these tensions are based on the assumption that they reflect irreconcilable positions, then one should expect that the incompatibility between AUSTERE and INDULGENT attitudes with respect to those tensions is justified. Well, the view about a correct approach to natural kindness in the

current study of the mental that I want to put forward is based on the consideration that such incompatibility is unjustified. A look at some of the advantages associated with these two perspectives can help introduce the rationale for my proposal.

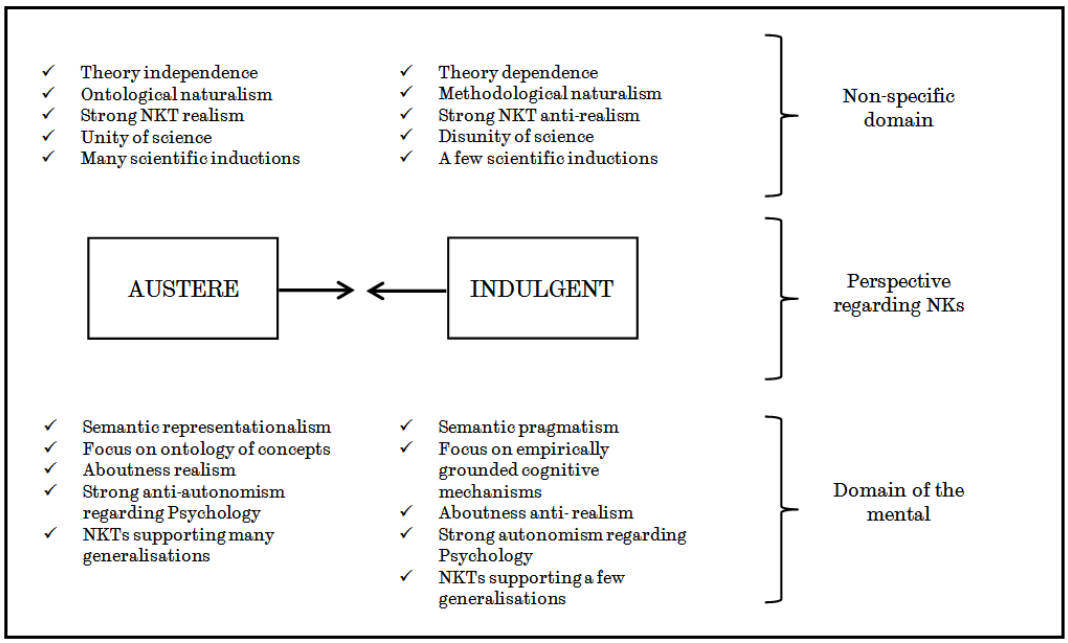

Diagram 1

AUSTERE and INDULGENT positions regarding natural kindness, as I have characterised them here, can be associated with a number of comparative explanatory advantages. I will conclude this chapter by arguing for the need to take both perspectives equally seriously. Later on, in Chapter 6, I will provide a positive view regarding the scientific relevance of concepts that is intended to subsume the apparently independent explanatory advantages of AUSTERE and INDULGENT positions. The main motivation for this positive view will be an attempt to render the theoretically unproductive tensions between those positions theoretically productive.

4. AUSTERE or INDULGENT? A tension to be reappraised

I have characterised the sources of an apparent tension underlying two broad opposing philosophical perspectives in the discussion of natural kinds. Similar

attitudes towards natural kindness can be recognised in the study of concepts, attitudes which, in some particular respect or another, some philosophers have acknowledged by distinguishing between two theoretical projects, namely one that is primarily the concern of philosophical theories of concepts vis-à-vis another that is primarily the concern of psychological theories of concepts. For instance, in discussing categorisation as one of the phenomena he thinks an adequate theory of concepts should explain, Prinz (2002) points out that:

> Philosophers rarely try to accommodate such psychological findings when developing their theories. In fact, some philosophers think that a theory of concepts need not explain categorization at all. The constituents of thoughts, they contend, may have little to do with the mechanisms by which we classify objects under the conditions psychologists explore. (p. 10)

In Prinz's view, excluding categorisation as an explanatory desideratum would introduce an unfair bias against psychological theories of concepts. Presumably, by including categorisation as one of the explanatory targets of a theory of concepts, the bias would be against philosophical theories that do not consider concepts relevant for the explanation of the mechanisms psychologists think are relevant for categorisation.

Machery (2009) has gone on to claim that psychological and philosophical theories of concepts are two fully independent projects. In his view, psychologists should simply focus on certain psychological structures that individuals use in higher cognitive processes and reject the use of the term 'concept' for scientific purposes. This radical conclusion illustrates the effect of choosing sides with respect to certain aspects of AUSTERE and INDULGENT.

As I will make it clear in the next chapter, the unproductive tension between AUSTERE and INDULGENT in the study of concepts should be re-interpreted in a different light, on the grounds that both perspectives address important concerns regarding the probable role of certain putative natural kinds in our mental life. What I maintain is that this tension is not resolved by simply choosing one of the two options (in one or more of the particular aspects of this tension listed in diagram 1), but by developing an adequate solution in which this unproductive tension is rendered productive. An alternative methodological perspective regarding an adequate

understanding of natural kindness in the study of the mental should help to dissipate many disputes between AUSTERE advocates and INDULGENT advocates.

There are certain requirements for the development of this alternative. A desired solution, as I will intend to show, rides on the acceptance that different theoretical frameworks, which are not clearly compatible with one another, can perfectly co-exist within the context of an immature scientific field. In this context, the use of vague or imprecise theoretical terms is compatible with a stricter form of scientific eliminativism in more mature sciences, and the problem of finding commonalities among psychological structures that might be involved in certain higher cognitive processes does not have to preclude the project of theorising about the nature of concepts. More importantly, the argument that a uniform notion of concepts should be replaced by a heterogeneous set of complex psychological structures on the grounds that these structures, but not concepts, pick out natural kinds can be seriously debunked. One reason for this is that, once the AUSTERE and INDULGENT distinction is reconsidered as a productive tension, it is perfectly acceptable for cognitive scientists to study different types of complex mental representations playing an important role in certain high level cognitive processes and, at the same time, study some type of less complex (or even, atomic) mental representations required as the constituents of thoughts *qua* complex mental representations, without claiming that the natural kind status of the former is a reason to question that of the latter.

5. Conclusion

This chapter addressed the problem of natural kindness within philosophy of science, with a focus on the study of concepts. I first presented some prominent philosophical accounts of the notion of natural kinds and the problem special sciences impose on a unified construal of that notion. After that, I characterised the possible sources of a distinction between two apparently opposing perspectives to natural kinds, namely AUSTERE and INDULGENT, and claimed that this distinction permeates the study of concepts in an unproductive way. In view of that, I argued for the need of a mediating methodological solution that overcomes the unhelpful tensions between the two perspectives. I maintained that these tensions demand some

important reappraisal and that this reappraisal should allow for a more productive perspective towards the notion of natural kinds in the study of concepts, namely one in which the advantages typically associated to one perspective are not viewed as a challenge for the other. Finally, I set forth some general requirements I think an adequate alternative solution should satisfy. These requirements will be instrumental in the proposal about natural kinds within the scientific study of mind next chapter will intend to sketch.

Chapter 6

‘Concept’ is a Legitimate Natural-Kind Term

1. Introduction

The aim of this conclusion chapter is to defend the thesis that the term ‘concept’ is a legitimate natural-kind term in the context of an immature science of the mind. To do so, I will sketch a positive view regarding theoretical terms in psychology relevant for the task of identifying putative natural kinds and lawlike generalisations about them. Contrary to those who are skeptical about the theoretical credentials of the concept of a concept, I maintain that a suitable approach to natural kinds in psychology must allow for the co-existence of different models of concepts and attitudes towards the study of concepts. In this sense, the main contribution of the current work is that the view of scientifically relevant terms I favour does not only object to concept eliminativism, but it also offers common methodological space for issues raised by otherwise irreconcilable positions within philosophy of cognitive science, as those motivated by the several tensions characterised in terms of the AUSTERE and INDULGENT distinction in Chapter 5. In the proposed view, those positions can productively fall under the qualified application of a common conception of natural kinds, without disregarding their diverse motivations and explanatory interests.

My main focus of this chapter will be on the claim that, once the tensions between AUSTERE and INDULGENT attitudes towards natural kinds (hereafter, simply AUSTERE and INDULGENT) are constructively reframed within the study of concepts, the heuristic advantages of applying a single common conception of natural kindness in two complementary ways (motivated by AUSTERE and INDULGENT, respectively) outweighs the prospects of a taxonomically relevant vocabulary constrained by AUSTERE and INDULGENT as two irreconcilable positions. In view of that, I will argue that a qualified application Homeostatic Property Clusters (HPCs) is—until further notice—a proper notion of natural kinds that can help distinguish concepts as putative natural kinds.

In short, the central argument of this chapter will be that, just as it is perfectly justified that an appropriate account of natural-kind terms in special sciences (notably, highly immature sciences) indulge in certain significant latitude *vis-à-vis* an appropriate account of natural-kind terms in more basic sciences (i.e., less immature sciences), it is perfectly justified that the theoretical terms within a highly immature special science such as psychology can be distinguished along the same lines. Accordingly, I will defend a combined perspective on 'concept' as a natural-kind term whereby the notion of HPC can be heuristically applied both as per AUSTERE and INDULGENT in order to identify putative natural kinds designated by the term 'concept'.

2. AUSTERE and INDULGENT: Together but not scrambled

In this section, I will argue, firstly, that different types of mental representations can be broadly characterised as HPC kinds in one of two different ways, at different degrees of complexity, depending on the type of causal mechanisms that sustain the property clusters that distinguish those two kinds. I maintain that the most prominent models of concepts within Cognitive Science can be sorted out according to this distinction. Secondly, I will argue that the traditional view according to which natural kinds share common properties must be taken with a pinch of salt when it comes to immature sciences. The reason for this is the unlikelihood that so-called reliable empirical procedures within highly immature (and loosely connected) sciences of cognition can (as yet) supply accurate objective descriptions of reality. In effect, given the multiplicity of theories and the ensuing methodological diversity in the study of concepts, it is hard to accept that everything one can say about concepts (and their nature) is contained in the explanations these sciences are in a position to provide. This latter claim is not meant to object to scientific realism as a positive epistemic attitude but, instead, it is meant to object to any strict application of a single conception of natural kinds within current Cognitive Science, simply because we do not as yet know enough about the kinds psychological generalisations should subsume. Thirdly, I will sketch a view of natural kinds where AUSTERE and INDULGENT can find common ground and I will argue for the methodological and heuristic advantages of this view. And, finally, I will provide a possible (and yet

unavoidably tendentious) reinterpretation of Machery's heterogeneous BoKs that does not challenge the scientific relevance of the term 'concept'.

2.1. Lower-MRs and Upper-MRs mental representations as two types of HPC kinds

I find it is fairly uncontroversial to claim that the probable existence of putative natural kinds which are highly complex (or molecular) mental representations is not incompatible with there being putative natural kinds which are minimally complex (or even non-molecular) mental representations. Call these two types of psychological kinds *Upper* and *Lower mental representations* (hereafter, Upper-MRs and Lower-MRs), respectively. For example, suppose prototypical representations of a given category (e.g., the prototype of an elephant) are highly complex mental representations that encode a set of statistical relations to minimally complex mental representations called the features of the prototype (say, e.g., lexical representations of tusks, trunks, etc.). In this case, as it might be, prototypes could be putative natural kinds at the level of Upper-MRs, and their features could be putative natural kinds at the level of Lower-MRs.

Another example might help to make the point clearer. Suppose, for the sake of illustration, that the content of many linguistic expressions is actually determined by the content of language-like mental representations. And suppose also that, as an empirical fact, it is common for language users (e.g., English language users) to talk about two different types of referents (say, discrete entities and events, respectively) by using two different types of linguistic expressions, say, nominal phrases (NPs), such as, e.g., "Edison", "the motion picture camera", "the invention of the motion picture camera", etc., and sentences (Se), such as, e.g., "Edison invented the motion picture camera", respectively). Now suppose that, just as it is possible to account for the complexity of linguistic expressions by means of syntactic transformations (e.g., Ses resulting from the combination of NPs and the like), so can the complexity of mental representations be accounted for in terms of the transformation of different types of canonically structured mental representations. Thus, whereas certain types of complex mental representations can be the bearers of the content expressed by certain Ses, certain other less complex mental representations could be the bearers of the

content expressed by certain NPs. In this case, as it might be, while those complex mental representations could amount to putative natural kinds at the level of Upper-MRs, the other less complex type of mental representations could amount to putative natural kinds at the level of Lower-MRs.

Most concept theorists have typically identified concepts with some form of Upper-MRs (e.g., prototypes, theories, exemplars, etc.) and a few of them have claimed that (some or most lexical) concepts are Lower-MRs (e.g., lexical atoms in a language of thought).[64] Thus, it is perfectly possible that, contrary to Machery (2009), the consideration that different prominent models of Upper-MRs fail to pick out a single homogeneous natural kind does not really challenge the view that there are concepts, since the term 'concept' could very well turn out to pick out either some other type of Upper-MRs or some type of Lower-MRs.

Machery (2009, 2010) has offered what might be considered the most explicit and straightforward eliminativist proposal regarding concepts (see Chapters 3 and 4) so far. According to this view, the theoretical term 'concept' should be replaced by three alternative models of psychological structures (i.e., prototypes, exemplars and theories) because 'concept', as opposed to each of his proposed models, does not pick out a single Homeostatic Property Cluster (HPC). The view of proper natural kinds qua HPCs is arguably the best available contribution regarding natural kinds within philosophy of science (see Chapter 5, section 2.2). The key feature of HPCs is the idea that the properties which are grouped together in the property cluster distinguishing a given natural kind are not accidentally grouped together, but due to, at least, one sustaining causal mechanism. However, since there is not just one way in which a given set of mental properties can be causally sustained, it is possible to object to the eliminativist conclusion of Machery's argument on the grounds that it overlooks other possible applications of Boyd's conception of natural kinds.

Indeed, since the question about what homeostatic mechanisms are relevant for which HPC kind is an *a posteriori* question, the issue of whether concepts form a proper natural kind is otiose. The term 'concept' is a term of art not so much because it is assumed to pick out a proper natural kind (as though it was the term in the vocabulary of a true theory of the world), but because it allows specialists to

[64] See Chapters 2 (section 2) and 3 (section 3.2), respectively.

problematize, theorise and provide a domain for certain mental phenomena and processes that are in need of explanation.[65] Likewise, the same consideration applies to the notion of HPC as a characterisation of *proper natural kinds*, since empirical generalisations about some putative natural kind do not *ipso facto* entail genuine natural laws about proper natural kinds. As de Sousa (1984) puts it,

> Virtually any kind can be termed 'natural' relative to some set of interests and epistemic priorities. Science determines those priorities at any particular stage in its progress, and what kinds are most 'natural' in that sense is always a real and lively scientific question. (p. 562)

For this type of reason, Machery's scientific eliminativism regarding concepts is eliminativism regarding the theoretical term 'concept' (i.e., a methodological or taxonomical issue), not eliminativism regarding a hypothesised putative natural kind (i.e., a metaphysical issue). Machery's claim that the class of concepts is not a homogenous class is really a claim on whether the terms 'prototype', 'exemplar' and 'theory' are all legitimate natural kind terms in psychology (i.e., an attempt to show that these terms can be said to pick out certain HPCs, the best *currently available* conception of natural kinds). In his view, those three theoretical terms should replace the notion of concepts because the term 'concept' is methodologically inadequate—the assumption being that the term 'concept' fails to pick out a homogenous HPC because none of the three terms 'prototype', 'exemplar' and 'theory' succeeds in individually doing so in a way that makes the other two terms explanatorily redundant. Here, the notion of HPC is playing a regulatory role for the acceptance of the claim that each of those three notions is a legitimate natural kind term in its own right. However, there is no obvious connection between the acceptance of this claim and the acceptance of the claim that the notion of concept fails to pick out a HPC.

Indeed, if someone thinks that prototypes, exemplars and theories are bad models of concepts, irrespective of whether they designate HPCs, then the realisation

[65] In this respect, for example, the very notions of prototypes, exemplars and theories as genuine theoretical terms in the study of cognition can be said to be the result of the attempt to develop an adequate and comprehensive account of concepts. In some way, the productive empirical study of prototypes, exemplars and theories is a by-product of the investigation of concepts and the kind of phenomena where concepts are typically assumed to play an explanatory role.

that Machery's proposed BoKs amount to distinct HPCs may be deemed an additional reason for rejecting them as a theoretical alternative to the term concept. The reason for this is that, being different natural-kind terms (i.e., kind terms that designate distinct putative natural kinds and support inductive generalisations independently from one another), none of them is able to subsume the others as competing models of concepts. Alternatively, what I suggest is a less drastic application of the regulatory notion of HPC. In particular, this application should acknowledge that, even though current psychological theories of concepts are probably wrong about the kind of proper natural kind the term 'concept' might actually designate, this latter term is a fruitful vehicle of inquiry in the empirical study of the mind.

I agree that Machery's proposed BoKs (i.e., the kind of putative natural kinds the psychological terms 'prototype', 'exemplar' and 'theory' are supposed to designate) can be specified in terms of HPCs, but, as I have argued in Chapters 3 and 4, I disagree that Machery's view can satisfactorily replace the role concepts are normally expected to play in a theory of cognition. Alternatively, what I claim is that the argument according to which the terms 'prototype', 'exemplar' and 'theory' designate putative natural kinds because they pick out HPCs is not incompatible with the view that concepts are natural kinds *even if those three Upper-MRs form HPCs*. The reason for this is that the putative referents of 'concept' may still be HPCs, even if they are not exactly the same as the HPCs which are the kinds of Machery's BoKs. Thus, as I shall now go on to argue, by distinguishing two different sources of probable homeostatic mechanisms, the notion of HPCs could be legitimately applied to candidates for putative natural kinds designated by the term 'concept' at the level of both Upper-MRs and Lower-MRs.

I propose to start by considering a distinction between computation and information processing. Even though the notions of computation and information can be interpreted in different ways (e.g., Piccinini 2015), cognitive scientists tend to conflate these two notions as though they were one and the same thing. As Piccinini and Scarantino (2010) puts it:

> Right around the time they entered psychology and neuroscience, the notions of computation and information merged into what seemed an appealing synthesis. Roughly, the mind/brain was seen as a computer, whose function is to receive information from the environment, process it, and use it to control the body and perform intelligent actions. […] Since then, computation and information processing have become almost inseparable—and often indistinguishable—in much literature on the mind and brain. (p. 238)

The main motivation for the merging of these two notions in mainstream Cognitive Science is the project of accounting for our characteristically inferential mental processes in terms of the causal transformation of complex or structured mental representations. In this context, mental representations are normally conceived of as vehicles of information and, in turn, the information those vehicles carry is conceived of as the semantic content of our mental representations. However, because there is not just one way of conceiving of the notion of information as a way to characterise the content the mental representations might carry (e.g., consider the distinction between referentialist and intensionalist accounts of mental content discussed in Chapter 4), one can very well distinguish between computational properties of our mental representations and their informational properties. Indeed, different authors have emphasised one or the other when characterising the causal link between certain psychological states (or processes) and our intelligent behaviour (or else, our higher cognitive competences)[66].

In particular, one could distinguish between information processes that determine the content of (Lower-/Upper-) MRs and the computational processes that are sensitive to the information-preserving structure of (Lower-/Upper-) MRs. In other words, one could distinguish between two independent types of causal mechanisms sustaining both Lower-MRs and Upper-MRs.

Suppose, for instance, some Lower-MR (say, CAT) is individuated by their referent (more specifically, by a certain set of properties tokened by the members of the category of cats) as per, say, some form of informational semantics (see related discussion in Chapter 4, section 6). In this case, CAT is an HPC kind which is

[66] For example, Stich (1983) has argued that psychological explanation could be couched in terms of a Syntactic Theory of the Mind, according to which mental states are identified with purely syntactic mental representations causally interacting with one another in virtue of their formal properties. Piccinini and Scarantino (2010), in turn, highlight the centrality of the notion of information in the sciences of the mind and the explanation of cognition (among other disciplines).

informationally sustained.[67] Likewise, it is perfectly possible that, along with the occurrence of CAT (e.g., alongside the causal tokening of CAT), certain related Upper-MRs are inferentially activated (e.g., Upper-MRs expressing that cats have sandpaper tongues, retractable claws and can see in the dark; that when turned upside down and dropped from a height, cats tend to land on their feet; that Bernardo likes cats; that some cats are like Garfield, etc.). Since inferences can be explained in terms of causal, information-preserving computation processes, Upper-MRs which are thus tokened can be interpreted as computationally sustained HPCs. Whether these related Upper-MRs are constitutive of the type of mental representation CAT is a matter of empirical investigation and philosophical dispute, but it is safe to say that they may be involved in certain higher cognitive competences such as categorisation (see section 2.4 for an elaboration of this latter point).

Thus, Lower-MRs can be proper natural kinds by the same standards that Upper-MRs can, a consideration that does justice to the importance different cognitive scientists have attributed to different types of mental representations in the study of concepts. For instance, while informationally sustained HPCs could support theorising about a possible naturalised account of how a causal mind-world relation can be secured, computationally sustained HPCs (e.g., prototypes) could support the explanation of crucial mental phenomena, such as for example, how it is that certain putative primitive concepts are learnt (e.g., Margolis and Laurence 2011).

2.2. Rendering a destructive tension constructive

In Chapter 5, I concluded that the study of concepts would benefit from a perspective towards natural kinds that renders certain unproductive tensions characterised as AUSTERE and INDULGENT productive. As I maintained there, an adequate empirical study of concepts requires abandoning the idea that adopting a given epistemological position motivated by AUSTERE (e.g., the view that natural-

[67] It would appear that, by characterising certain (Lower-/Upper-) MRs in terms of HPCs, strict essentialism about putative natural kinds is ruled out. But this is exactly the kind of thinking my proposal wants to object to, since it assumes an unproductive methodological tension between AUSTERE and INDULGENT. Alternatively, since how it is that sustaining causal mechanisms actually work is an *a posteriori* question, one can leave the related question of whether or not informational semantics involves strict essentialism open. For present purposes, the notion of HPC is compatible with strict essentialism on pragmatic grounds. If strict essentialism is true, property clusters are a good heuristic tool to theorise about putative natural kinds in the study of concepts.

kind terms in psychology should support many inductive generalisations) excludes the scientific relevance an apparent antagonistic position motivated by INDULGENT (e.g., the view that natural-kind terms in psychology do not need to support many inductive generalisations), and vice versa.

'Austere' and 'Indulgent' are two labels I have used to characterise opposing attitudes towards a number of focal yet unresolved issues within both philosophy of cognitive science and, more specifically, the study of concepts. Because advocates of AUSTERE and INDULGENT (with respect to each of those unresolved issues) have made substantial cases in defence of each of their preferred positions, the possibility of bringing their contribution within a common theoretical framework, as I want to argue, is a perfectly coherent option. What follows is an attempt to sketch this option.

In Chapter 5, I refrained from proposing a general criterion to distinguish between all the different tensions depicting AUSTERE and INDULGENT attitudes, since being committed to one of these attitudes with respect to a given issue does not entail being committed the same attitude with respect to a different issue. Notwithstanding that consideration, in this Chapter, I want to go on to claim that, if there is one feature that may be said to pervade most (if not all) of the several positions labelled as AUSTERE and INDULGENT, I believe it is accommodating the metaphysical and methodological understanding of natural kinds. According to the metaphysical conception, natural kinds are objective (or theory independent), homogenous, and their members share a common essence. Additionally, from the perspective of this metaphysical conception, natural kind terms are to be understood in terms of a strong natural-kind-term realism, that is, as though they designated the kinds which are relevant to the laws of a completed special science (given the charitable assumption that the current state of a science is a stage in a process that leads to such a completed science).[68] By contrast, according to the methodological conception, natural kinds are theory-dependent, heterogeneous, and their members do not share common essences. Moreover, methodological natural kind terms are

[68] This way of depicting natural-kind terms is equivalent to what de Sousa (1984) dubs as 'rigid designators', namely, "[natural-kind terms that] pick out members of the same class or lumps of the same stuff in every possible world […] even if we are ignorant or mistaken about what that class or stuff really is." (p. 563-564)

committed to a strong natural-kind-term anti-realism, in that they are not expected to designate objective natural kinds.

An important feature of the metaphysical conception is that it allows for objectivity to be construed in terms of the independence of Ontology from Epistemology, that is, in terms of the independence of the kinds of things that there actually are in the world (whatever they might be) from the epistemic mechanisms that allow us to access them. Prominent philosophers (e.g., Rey 1983, 1985; Fodor 1998) have argued that overlooking the independence of Ontology from Epistemology has been the source of much confusion in the study of concepts. In particular, as these authors claim, one important reason for this confusion has to do with a tendency to conceive of the essential properties necessary for the individuation of some type of thing in the world as being dependent on the mental capacities that enable us to access them.

In turn, one prominent characteristic of the methodological conception of natural kinds rides on a certain degree of scepticism towards the very independence I have just pointed out. Indeed, if the independence of Ontology from Epistemology is thought to be irrelevant for scientific purposes, then the notion of a natural kind, as Dupré (2002) suggests, may be just deemed as "a (more or less) [sic] useful methodological concept" (p. 31). In this case, objectivity is a mere presumption and scientists should seek only empirical adequacy for their theories (van Fraassen, 1980). A focus on empirical adequacy has proved to be compatible with the use of theoretical terms that do not seem to clearly carve nature at its joints (i.e., terms that do not clearly pick out putative objective natural kinds). Cowie (2009) provides a good example with respect to the productive use of the term 'innate' in different areas of enquiry such as language acquisition and developmental sciences. Here, a most agreeable remark is that, despite the multiplicity of meanings theorists have attributed to the term 'innate', this term has played a significant role in motivating serious discussions among biologists and psychologists. Thus, in Chapter 1 (sections 4 and 5), I proposed a taxonomy of eliminativist arguments and claims where certain types of eliminativist projects rejected the existence of a given type of thing in the world without abandoning the theoretical terms normally used to designate those ontologically suspect referents. In the light of these considerations, I concur with the

view that vague or imprecise terms can undeniably make some heuristic contribution in any developing scientific field.

I agree that objectivity is no more than a presumption, especially in the context of immature special sciences where the use of vague theoretical terms is characteristic and probably unavoidable. In other words, these sciences need to assume certain kinds as natural, but it is understood that such assumptions are defeasible and corrigible in the course of scientific inquiry. But I disagree that a mere methodological conception of natural kinds can eventually uncover proper natural kinds: there has to be more to the existence of a genuine natural kind than pragmatic convenience, even the pragmatic convenience of the scientific community. The main reason one can provide for supporting this claim is de Sousa's consideration that the taxonomical success of a given science exclusively on the basis of empirical adequacy alone is both irrelevant for individuating objective natural kinds (kinds which are so intrinsically) and even compatible with the kind of pragmatism Locke advocated. In Locke's empiricist view, as de Sousa (1984) points out, what exists is unknowable and our concepts for natural kinds are ontologically arbitrary. Supposedly, it is the boundaries of these concepts that allow us to establish the natural kind a particular thing belongs to, but since the concepts are framed on a contingent basis and their extension is determined *intensionally*, they do not provide insight into the nature of objective natural kinds. In de Sousa's words, "[those concepts] at best represent pragmatic convenience." (p. 563).

My suggestion is that an appropriate methodological view of natural kinds should go beyond the alleged opposition between metaphysical and methodological matters. By going beyond that tension I mean to say that the methodological conception of natural kinds should also incorporate challenges posed by metaphysicians as part of its heuristic role in special sciences. For example, at the beginning of this section I mentioned my scepticism about the capability of current empirical theories of concepts for providing a satisfactory answer to the question of what concepts are. Suppose then that some sort of traditional ontological analysis dealing with the problem of the nature of concepts is able to offer certain *transcendental arguments* for or against some explanatory strategy in the study of cognition (cf. Fodor 1998). While such arguments can be deemed scientifically irrelevant due to their typically non-empirical character, there does not seem to be an

in-principle reason for opposing their contribution as *additional constraints* on argument to the best explanation in an emerging scientific field where initial progress may tend to resist unification.[69] For this reason, not only do I favour the idea of embracing the tension between AUSTERE and INDULGENT as a positive tool of progress, but I regard it as a requirement for correct theorising about natural kinds within the context of immature special sciences. I incline to believe that the more immature a given science is, the more it is that philosophers of science should learn how to live with this tension (i.e., the more they need to acknowledge it as a positive and productive tool).

More specifically, what I propose is a focus on a constructive reading of the tension in question. It is common that, in developing theories of concepts, their proponents take a certain *destructive attitude* towards some model of concept or another. There is nothing really wrong about these attitudes per se unless they are insufficiently justified. For example, if someone thinks she has good reasons to believe that concepts must be identified with some type of unanalysable atomic mental representations, then she may be justified in adopting a destructive attitude towards models of analysable concepts that directly challenge the thesis that concepts are atoms. Here, by adopting a justified destructive attitude I mean to say that someone has (what she thinks are) good reasons to conclude both that their opponents are wrong about the kind of things concepts might actually be and that their relevant theories have to be rejected.

I claim there aren't good reasons to make the case that, for theories and models of concepts advocating either AUSTERE or INDULGENT (in one particular respect or another), a destructive attitude towards theories and models advocating their opponent perspective should be preferred over a constructive one. On the contrary, given the significant variety of empirical findings within the psychology of concepts (see, e.g., Murphy 2002, Machery 2009, c.4) and the different philosophical views about the possible nature of concepts (see, e.g., Margolis and Laurence 2007; Prinz and Clark 2004) that have motivated the development of highly dissimilar theories of concepts (e.g., Fodor 1998; Peacocke 2002; Prinz 2002; Hampton 2006), a

[69] Of course, including this type of contribution as an additional argument for the best explanation need not entail any strong commitment to some sort foundational role of ontology with respect of psychology.

constructive attitude on the part of theories endorsing AUSTERE vis-à-vis theories endorsing INDULGENT (or, *mutatis mutandis*, theories endorsing INDULGNET vis-à-vis theories endorsing AUSTERE) should be preferred. Destructive attitudes, as characterised above, can be kept to aspects of theory construction and evaluation that do not appeal to a particular preference for AUSTERE or INDULGENT, as though the adherence to one of these perspectives was relevant for establishing a comparative advantage. Note that this claim is compatible with adopting destructive attitudes towards theories and models within the context of more established and less immature scientific areas of enquiry, irrespective of whether AUSTERE or INDULGENT is endorsed. My view is that a constructive reading of the tension between these two perspectives is a more sensible option when it comes to the *current* study of concepts.

There are a number of issues with respect to which the study of concepts would benefit from acknowledging the need to reappraise tensions constrained by AUSTERE and INDULGENT in a different light. Since, as I have argued, there isn't as yet any viable proposal to seriously challenge the scientific relevance of the term 'concept', cognitive scientists should welcome methodological alternatives that allow for the proliferation of relevant hypotheses, especially in a domain of inquiry where abandoning certain elusive theoretical terms might be a hasty decision. In the following section, I will recommend that the notion of HPCs be applied for two different yet compatible heuristic purposes. Additionally, I will illustrate the sort of issues that this model for the application of HPCs could help to combine in one single domain of enquiry.

2.3. A qualified application of HPCs in the study of concepts

Within cognitive science, enquiry into concepts and their role in our mental life proceeds as if we know both that there are concepts and that we can distinguish many of them. It then goes on to identify the properties that are typical of concepts, properties that should support inductive inferences about the things that fall under the extension of the concept. Concepts are typically identified with the tokening of mental representations and the properties that different competing theories deem as typical of concepts vary from one another, depending on both the main role concepts are supposed to play in our mental processes and how they are to be individuated.

For example, as described in Chapter 3, conceptual atomism (i.e., the view that concepts are unstructured) conceives of concepts as the basic constituents of thoughts (e.g., in a language-like representational medium) and aims at formulating conditions for the individuation of concepts (i.e., identity, semantic and possession conditions) without appealing to epistemic factors. Conceptual atomism is concerned with the properties of concepts that are metaphysically necessary for them to be what they are. By contrast, all the main psychological theories of concepts described in Chapter 2 are theories about the kind of structure that certain mental representations must have such that they can account for processes involved in many of our higher cognitive competences (e.g., categorisation, inference, etc.). These theories are not concerned with essential properties of concepts but the conditions that agents like us must satisfy to possess concepts. These conditions are typically specified in terms of epistemic factors such as certain inferential or recognitional abilities.

When categorising theories of concepts which identify concepts with mental representations, it is common to distinguish between primitive (or unstructured) and complex (or structured) concepts (Laurence and Margolis 1999, pp. 4-5). I have made a similar distinction between putative Lower-MRs and Upper-MRs that are HPCs, the difference being that the category of Lower-MRs that are HPCs is slightly permissive as to whether our most basic mental representations are atomic or minimally structured. The reason for this is simply the acknowledgement that there is no agreement about the nature of the most basic component of our mental representations (e.g., whether or not they are sensory or perceptual in character). This degree of permissiveness even allows us to read this distinction as a relative one. A given mental representation *M* might be Lower-MR with respect to one mental representation *N*, but Upper-MR with respect to another mental representation, *P* (such as, e.g., when CAT is deemed to be a molecular Lower-MR by the standards of, say, some statistically-based account of concepts and non-molecular Lower-MR by the standards of an atomistic account). The point to emphasise here is that one can find the distinction between Lower-MRs and Upper-MRs compelling irrespective of whether or not one is a relativist in this respect—I hope it will become clear later on that, in the context of an immature cognitive science, it is a good thing that this is the case.

I have also distinguished between HPCs which are informationally and computationally sustained. Mainly for heuristic purposes, I propose to apply each of the two versions of causally sustained HPCs to any Lower-MRs and Upper-MRs for which the term 'concept' is a possible designator (see diagram 2). To regiment this application of HPCs, the following qualification is in order: while computationally sustained HPCs should be preferred where the goal of enquiry is empirical adequacy, informationally sustained HPCs should be preferred where the goal of enquiry is the metaphysical individuation of essential properties.

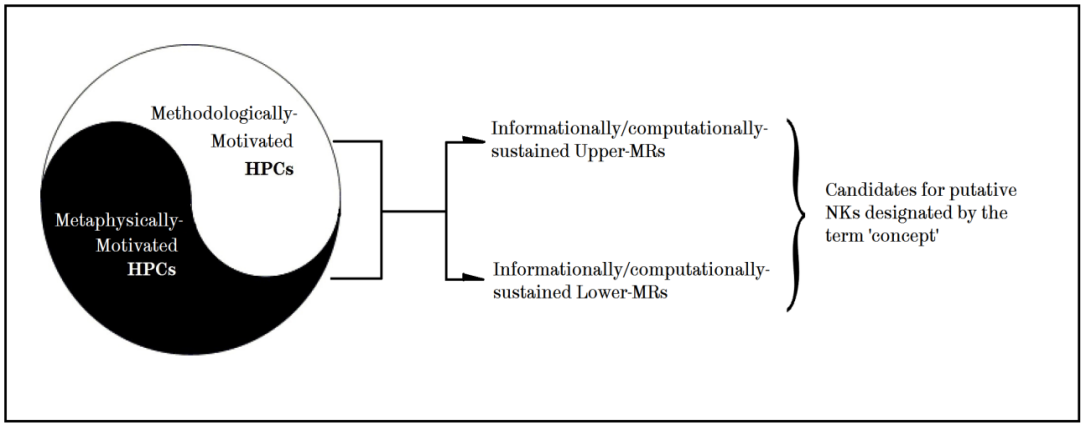

Diagram 2

Once the tension between AUSTERE and INDULGENT positions is accepted as a positive investigative tool, important controversies can be said to have a place in a common domain of enquiry where progress requires a sort of 'vulgar pragmatism' regarding our theoretical vocabulary. Controversies about the role of natural kinds in theory construction and evaluation are not an exception.

According to Machery (2010), "The study of concepts is in an odd state of disarray" (p. 195). The reason for this, he goes on to say, is that current theories of concepts have failed to provide a general theoretical framework for all the known phenomena in need of explanation and there is little agreement as to what type of thing (i.e., what type of putative natural kind) the term 'concept' might actually designate. However, for all kinds of reasons, the cognitive science community has

objected to the elimination of this term.[70] Thus, since this very disarray may be a good reason to be cautious about the premature elimination of a widely used term, the eliminativist's claim that the class of concepts is too heterogeneous to form a natural kind is pointing at something that does not have to be deemed an obstacle to scientific progress. Again, for want of a more precise theoretical terminology, philosophers and cognitive scientists can't avoid using the term 'concept', a term that has helped to raise and discuss issues that are relevant for both philosophers and psychologists. The following list of relatively familiar controversies can help to illustrate the relevance of keeping a constructive understanding of positions typically framed in terms of AUSTERE and INDULGENT:

(i) *Discussion of concept possession*. Fodor (2004) has distinguished between theories which focus on concept possession from a *pragmatist orientation* (e.g. having concept C depends on agents' capacity to distinguish concept C from non-C or her capacity to recognise the validity of some C-involving inferences) and theories which focus concept possession from a *Cartesian orientation* (e.g., having C is being able to think about concepts 'as such'). Fodor favours a theory of concepts that is consistent with the Cartesian orientation (viz., his version of Conceptual Atomism), since he thinks pragmatism about concept possession is wrong. What is wrong, he argues, has to do with concept theorists' assumptions about the correct order of explanation. His view is that an account of concept possession is parasitic on (and, therefore, not prior to) an account of concept individuation. From the point of view of those who think this view is probably true, a metaphysically-motivated application of HPCs can be said to serve the purposes of an adequate account of concept possession because, in the Cartesian view, concept possession is not an epistemic condition.

By contrast, as I mentioned earlier, there are prominent psychological and philosophical theories of concepts whose main concern is the study of the conditions that agents like us have to satisfy in order to possess concepts (e.g., see Peacocke 1992 for a philosophical theory; see Chapter 2, for psychological theories). For these theories, different epistemic capacities (e.g., inferring, recognizing, categorising, etc.)

[70] See, e.g., the open peer commentaries in Machery 2010.

are constitutive of what it is to possess concepts, and, as a result of that, the answer to the question of what concepts are is derived from the conditions for concept possession (i.e., a concept is a capacity, viz. that which the possession conditions of a concept are indicative of). From the point of view of those who think this alternative view is adequate for purposes of explanation, a methodologically-motivated application of HPCs can serve their explanatory purposes because, among other things, this conception of HPCs is based on the kind of cognitive mechanisms whose primacy the Cartesian view of concept possession rejects.

Concept pragmatism and concept Cartesianism are presented as two irreconcilable positions that can't be part of a common theoretical framework. However, both positions can help (and have helped) to raise important concerns relevant to the discussions of concept possession: can the investigation into the nature of concepts be a direct concern of empirical research? To what extent is the satisfaction of the conditions for concept possession a sufficient requirement for an account of the nature of concepts? Since the use a common terminology among theorists has been instrumental in raising these concerns in the study of concepts, it is reasonable to conclude that a suitable and constructive approach to the corresponding controversies should allow for both positions (i.e., pragmatism and Cartesianism) to co-exist.

(ii) *Disputes about the structure of concepts*. Current psychological theories of concepts tend to assume that concepts are a single, uniform type of mental representation, but there are concept theorists who think this assumption is unjustified (e.g., Weiskopf 2009; Machery 2009). The reason for this (or so the argument goes) is that none of the most prominent theories of concepts has been able to explain all the empirical data related to the kind of phenomena where concepts are supposed to play an important role. As a result of this, none of those theories can be clearly said to be in a better position to displace its competitors, a realisation some think is sufficient justification to claim that the assumption that concepts form a single, uniform kind of mental representation is seriously mistaken. A reaction to this impasse is the alternative view that some form of *concept pluralism* is in a better position to explain

the available data. According to this view, the class of concepts really comprises many different types of mental representations.

Psychological theories have identified concepts with different types of mental structures and each of the proposed models of concepts has shown to be more suitable to explain certain aspects of our cognitive mind than others (see Chapter 2, section 2.2). Hence, those theories have centred around the kind of structure that the mental representations involved in certain cognitive processes must have in order to account for certain higher cognitive competences (e.g., inference and categorisation). In this sense, debates about the structure of concepts (notably, whether or not concepts comprise a single, uniform type of mental representation) can be said to aim at satisfying the condition of empirical adequacy.

Prima facie, it would appear that a pluralist theory that is able to subsume all the most prominent models of concepts has a clear explanatory advantage over any competing theories. However, the argument for concept plurality may not be very persuasive for someone who thinks that empirical adequacy is not the most important criterion relevant for both theory construction and evaluation. For instance, competing theories of concepts could also be assessed in terms of the extent to which they satisfy some non-negotiable conditions (e.g., Fodor 1998), given some preliminary metaphysical commitments (e.g., a commitment to a general metaphysical view of the mind and its working). Here, objections to a given theory regarding its empirical adequacy are still relevant, but they do not need to result in the conclusion that concept pluralism must have an inherent advantage over other less permissive theories when it comes to discussions on the structure of concepts.

Suppose, for the sake of argument, that the *constraint of compositionality*[71] as an explanatory desideratum is deemed to be more important than empirical adequacy. Advocates of both views (i.e. uniformity and pluralism about the structure of concepts) have emphasised the importance of concepts in accounting for the productive and systematic character of thought (e.g., Fodor and Lepore 2002; Weiskopf 2009). They also agree that the characteristic productivity and systematicity

[71] Roughly, 'compositionality' refers to the principle according to which the meaning of a thought is inherited from the meaning of its constituents, together with their structural arrangement.

of thought are two arguments for the compositionality of our mental representations.[72] However, the compositionality constraint has been particularly instrumental in the defence of Conceptual Atomism (e.g., Fodor 1998), and yet that is a view that many cognitive scientists would deem to be a non-starter. Someone might reasonably wonder why the compositionality constraint has not received similar attention in regard to other views about what concepts are.

It should be noted that, strictly speaking, the productivity and systematicity of thought are a piece of evidence for the claim that thoughts are structured mental representations with concepts as their constituents, not for a claim about the structure of those constituent concepts. What this suggests is that the constraint of compositionality is relevant for any theory that conceives of concepts as the constituents of thoughts both in philosophy (e.g., Fodor 1998) and psychology (e.g., Carey 2009). In this sense, as it might be, the constraint of compositionality is relevant for practically all theories of concepts, irrespective of whether they identify concepts with structured or unstructured mental representations. Hence, someone could legitimately object that difficulties in accounting for the productivity and systematicity of thought are not exclusively grounds for treating models of structured concepts as implausible.

However, a characterisation of concepts as the constituents of thoughts is not shared by all concept theorists and this lack of agreement about this characterisation reflects radically different views about what a theory of concepts should achieve. Some philosophers (e.g., Prinz 2005; Edwards 2009) have pointed out an alleged incompatibility between views of concepts as arbitrary symbols primarily used for representing categories and views of concepts as the mechanisms that allow us to draw inferences, recognise categories, etc. Similarly, Machery (2010) has controversially distinguished between two fundamentally different theoretical projects associated with the term 'concept': the first project involves an "attempt to determine

[72] The term 'productivity' is used to refer to our cognitive capacity to entertain an unbounded number of thoughts. In turn, 'systematicity' is normally used to refer to two empirical facts: the first is that our capacity to entertain certain thoughts is intrinsically connected to our capacity to entertain certain others. The second is that our capacity to make certain inferences is intrinsically connected with our capacity to make certain others. Thus, the systematicity of our cognitive capacities would explain why it is unlikely that someone who is able to entertain the thought that, say, *politicians sometimes behave like celebrities* but unable to entertain the thought that *celebrities sometimes behave like politicians* (cf. our capacity to infer the thought that X from X&Y and our capacity to infer the thought Z from X&Y&Z).

the conditions under which people are able to have propositional attitudes about the objects of their attitudes" (p. 199) and the second involves an attempt to explain the mechanisms underwriting our abilities to categorise, draw inferences, etc. Given the way this distinction is characterised, it is unclear whether the notion of concepts as the constituent of thoughts is equally relevant for both theoretical projects.

Note that the first project can be said to be consistent with the view that concepts are the constituents of thoughts, since the generality of intentional explanations of people's behaviour [73] presupposes that the same intentional explanation can apply to different people because their mental states are composed of the same concepts (Prinz 2002, c. 1). By contrast, theories falling under the second theoretical project do not aim at the establishment of intentional generalisations and, hence, at an account of our cognitive mind where the view of concepts as thought constituents plays an important role. Instead, given their alternative explanatory targets, these theories tend to construe relations of constituency with respect to the posited structure of concepts (Prinz 2005). As Weiskopf (2010) has remarked,

> "Prototypes, exemplars, and theories are all complex representations that bear structural relations to their parts, over which inferences, similarity computations, and the like, might operate." (p. 228).

Interestingly, it is by considering the apparent incompatibility between theories with different explanatory targets that one can raise issues and worries about each of the two projects: how is it that theories that characterise concepts as the constituents of thoughts can accommodate available empirical findings (e.g., the role of prototypical representations in an account of typicality effects) supporting theories that characterise concepts as the mechanisms underwriting particular cognitive competences (e.g., categorisation)? How is it that theories of concepts aimed at empirical adequacy can respond to the characterisation of concepts as the constituents of thoughts, given their limited resources for satisfying the compositionality constraint? Does the explanatory relevance of constituency relations regarding the structure of concepts support the explanatory irrelevance of constituency relations

[73] That is, explanations of behaviour by appealing to the causal role of mental states.

regarding the structure of thoughts? It is fair to say that, while the jury is still out with respect to issues like these, it is because theorists count on a common way of talking about concepts that these concerns can be raised.

Radically different views about the goals of a theory of concepts can help explain the motivations underlying radically different views about the structure of concepts. Consider, again, concept pluralism and concept atomism as reactions to the limited explanatory scope of current psychological theories of concepts. Where the goals of a theory are motivated by the explanatory role of concepts in particular cognitive abilities (e.g., categorisation, inference, etc.), concept theorists can be justified in inferring that the class of concepts comprises many types of mental representations and that, perhaps, some form of concept pluralism is probably the best option capable of satisfying the condition of empirical adequacy. Alternatively, where the goals of a theory are motivated by a commitment to a certain broad metaphysical picture of the mind (e.g., that our mental processes take place in a system of representations with syntactic constituent structure), cognitive theorists can be justified in inferring that the class of concepts comprises a uniform type of mental representations and that, perhaps, some form of conceptual atomism is probably the best option capable of satisfying the compositionality constrain.

In this respect, I maintain that trying to find common ground between theories with fundamentally different motivations is to be preferred to the destructive option of favouring a myopic development of just one type of theory. More importantly, there are no good reasons to conclude that, because different views about the structure of concepts (notably, uniform and non-uniform types of representations) are motivated by radically different assumptions about explanatory goals of a theory of concept, the term ‘concept’ is a bad natural kind term.

Instead, what I suggest is that, when ‘concept’ is used as a theoretical term aimed at designating empirically-adequate psychological kinds (i.e., prototypes, exemplars and theories), a methodological application of HPCs should be favoured. The reason for this is that this application of HPCs is compatible with the operational relations of constituency that a complex mental representation is supposed to bear to its structural components (e.g., the statistical relation between a prototypical representation and the features that it might encode). Alternatively, a metaphysically-

motivated conception of HPCs is more useful when 'concept' is combined with less empirically-adequate hypotheses about concepts. The notion of informationally-sustained HPCs is a useful heuristic tool for exploring hypotheses about the compositional character of our thought processes.

(iii) *Psychological generalisations*. Admittedly, there is a kind of general agreement that philosophers and psychologists do not talk about exactly the same kind of thing when they use the term 'concept' (see Chapter 5, section 4). If we take a closer look at the particular theories of concepts psychologists and philosophers have developed, one might reasonably wonder if there any two theories about concepts that are really about the same kind of thing. For comparative purposes, it is common practice to provide idealised versions of theories that attempt to do justice to the variety of theoretical proposals available in the literature on concepts. As you would expect, without a common idea of what concepts are, the question of what a good theory of them can explain remains just as controversial as practically anything else theorists might want to say about concepts. In order to explore possible answers to this question, it can be useful to adopt a wider perspective and try to identify where there can be common ground regarding the scientific study of concepts.

Since almost every philosophical claim about concepts seems to be controversial (Peacocke 2009), it should come as no surprise that one of the widest agreements in the scientific study of concepts has nothing to do with concepts in particular: just as in practically any other special science (e.g., chemistry, biology, economics, etc.), cognitive scientists propose theories about certain particulars and assess those theories in the light of certain evidence typically collected by relatively standard procedures that the scientific community recognise as reliable. Similarly, as in many other sciences, cognitive scientists develop a theoretical vocabulary to theorise about the particulars they study and attempt to discover causally grounded generalisations on the assumption that the existence of those particulars is what justifies our scientific inferences. In this sense, the more a given particular under study successfully supports the discovery of scientific inductive generalisations, the more scientists would be justified in inferring that those particulars are candidates for natural kinds. What is special about the current state of cognitive science is the lack of

agreement about the putative nature of the mental particulars that may be involved in an explanation of cognition and, hence, what a good theory of these particulars should explain.

As you may recall from the relevant discussion on special sciences in Chapter 5 (section 2), the autonomy of psychology has been defended on the grounds that natural kind predicates in the laws of an ideally completed science of the mind are not reducible to the natural kind predicates in the laws of any other science. If this view is correct, and if psychology is to become that correct science of the mind, then it is reasonable to wonder about the aims of the generalisations that are interesting for psychology as an autonomous special science. One possible answer to this question stems from the programmatic consensus shared by the community of cognitive scientists committed to the classical symbolic/computational paradigm of cognition. According to this consensus, a science of mind should aim at the formulation of law-like generalisations that count as the basis for the explanation of intelligent behaviours, where intelligent behaviour is roughly understood as one which is guided by goals and purposes (Simon and Kaplan 1989).

At the heart of the classical paradigm of cognition is the assumption that the cognitive mind is an information-processing device that contains mental representations in virtue of whose structure information-processing is realised. Accordingly, if concepts are to play a role in cognition, then they must also be mental representations. Indeed, a general agreement within philosophy of cognitive science is that concepts are the basic constituents of the mental representations that are scientifically interesting for a science of cognition. In particular, concepts are assumed to be the primary bearers of the intentional contents that complex mental representations are the vehicles of. Hence the view that a theory of concepts is at the centre of a general theory of cognition.

The importance of identifying putative natural kinds is that they could tell us what the laws of a given special science should govern (Dupré 2002). Thus, if a scientific intentional psychology is to aim at the formulation of lawlike generalisations that help to account for our intelligent behaviour, then a psychology of concepts should aim at identifying the kind of things over which those generalisations might quantify. Usually, the generality of intentional explanations is said to be

grounded in the attribution of mental states in the form of propositional attitudes (e.g., O believes that P; O desires P). One reason for this is that propositional attitudes are thought to involve the tokening of mental representations that have concepts as their constituents. Another reason is the assumption that concepts are capable of being shared. Thus, the reason why the attribution of propositional attitudes as an explanation of behaviour (e.g., Claudia went to the box office because she desired to get a ticket for tonight's film and believed that she could buy one there) can generalise and subsume different people or the same person at different times (e.g., Constanza, Trinidad and Nico can all go to other box offices for the same reason that Claudia did) comes down to the fact that the concepts underwriting intentional generalizations must be shareable among different people and the same person at different times. For this reason, the shareability requirement is sometimes deemed as a core motivation for a theory of concepts (Edwards 2010).

To put it in more general terms, intentional generalisations in psychology generalize over content-bearing intentional states. If such generalizations are to be possible at all, then we must be able to represent intentional states which have the same content in a potentially endless number of different agents and situations. Our way of attributing these contents is by shared human concepts. Note that the same can't be done in the case of other species, precisely because we cannot justify the assumption of shared concepts. The intentional explanations we sometimes attribute to the behaviour of our pets (e.g., explanations of why a dog is pawing at the earth in attempt to get a bone) are typically anthropocentric, in that they really reflect the kind of explanations we would give of the behaviour of agents like us if they were acting like our pets normally do.

However, such motivation for a theory of concepts might not be relevant for a concept theorist who is not committed to the view that psychological generalisations involve intentional explanations. Similarly, as it might be, the shareability requirement may very well be acknowledged as a concern for a theory of concepts that is not aimed at supporting intentional generalisations (e.g., Carey 1985). Indeed, current theorising about concepts in psychology does not seem to be motivated by the programmatic consensus of classical Cognitive Science mentioned above. Or else, the motivation is something like a re-interpretation of that consensus. In this re-interpretation, psychological generalisations are about intelligent behaviour as long as

intelligent behaviour can be analysed in terms of some set of higher cognitive competences, notably categorisation. Indeed, to an important extent, the current psychology of concepts based on this re-interpretation of the consensus seems to straightforwardly reduce to the psychology of categorisation. As Prinz (2002) puts it, "psychologists typically postulate concepts to explain categorization" (p. 272). Machery (2010) goes on to characterise the current motivation for a theory of concepts in the following way:

> Why do cognitive scientists want a theory of concepts? Theories of concepts are meant to explain the properties of our cognitive competences. People categorize the way they do, they draw the inductions they do, and so on, because of the properties of the concepts they have. (p. 199)

The contrast between the different motivations for a theory of concepts I have described is probably best described in terms of a transition from a focus on intentional generalisations to a focus on models of categorisation. Intentional generalisations can be said to be aimed at a kind of 'hard' causally grounded generalisation, in that they are committed to the formulation of lawlike generalisations. By contrast, the causally-grounded psychological generalisations relevant for the development of models of categorisation are 'soft', in that their relevant casual factors are modified strictly in terms of what the available evidence dictates.

As might be expected, my suggestion is that the term 'concept' can apply to both metaphysically-motivated HPCs that are postulated to formulate 'hard' causally grounded generalisations and methodologically-motivated HPCs that are postulated to support 'soft' causally grounded generalisations.

2.4. Machery's heterogeneous BoKs re-interpreted

Finally, we are now in a position to suggest a relevant reinterpretation of the type of HPCs that the notions of prototype, exemplar and theory are supposed to pick out. In Machery's view, the claim that the taxonomical term 'concept' is empty

derives from the realisation that prototypes, exemplars and theories can be characterised as three distinct HPCs. In my view, the same realisation does not entail the claim that the term 'concept' is empty since those three HPCs are compatible with the probable existence of computationally sustained HPCs that are not concepts. I ground this view on the acceptance of two empirical claims. The first claim is that computationally sustained HPCs satisfy the traditional formality condition of the classical paradigm of Cognitive Science according to which our mental processes are defined over the structure of our mental representations. If this is correct, then it is plausible to think that they support some sort of "null hypothesis" according to which the view that there are prototypes, exemplars and theories does not challenge the existence of traditional syntactic representations.[74] The reason for this is that it is plausible that the information typically associated with those three constructs is related to the content, not the form, of already interpreted Upper-MRs (notably, thoughts about prototypical or particular dogs). The second claim involves a particular characterisation of the distinction between constitutive and background knowledge.

It is reasonable to think that, together with the tokening of some particular mental representation, certain concomitant activation of Upper-MR-based inferential processes takes place. Suppose concept C corresponds to some type of mental representation that we are using in certain cognitive processes in a given situation, such as, for example, when we communicate some C-involving thoughts to someone else in a given context. The suggestion is that we should interpret whatever context-independent information that concept C bears when we use it to communicate with others as knowledge that is constitutive of C. Alternatively, the context-dependent information about C contained in those Upper-MRs involved in concomitant inferential processes should be interpreted as background knowledge.[75] Given this characterisation of constitutive and background knowledge, I suggest that we identify prototypes, exemplars and theories with the type of computationally sustained HPCs that concomitant inferential processes bring to bear in the form of certain systematically organised background knowledge. For example, if CAT is a concept, then CAT-prototypes, CAT-exemplars and CAT-theories might correspond to

[74] This claim is motivated by the characterisation Pylyshyn (2002) provides of a "null hypothesis" with respect to imaged-based reasoning.
[75] Cf. Machery's alternative characterisation of the distinction between 'invariantism' and 'contextualism' (Machery 2015).

interpreted Upper-MRs encoding information about what we know (or believe) about cats.[76] The systematic role of these Upper-MRs in categorisation tasks as informed by empirical data suggests that they are legitimate psychological HPCs.

I have objected to the claim that these types of HPCs can replace the role concepts are expected to play in the explanation of cognition (see Chapters 3 and 4), but there is significant empirical evidence to support the claim that they play an important role in certain higher cognitive processes, notably categorisation. In this sense, I agree with Machery that the most prominent theories of concepts fail to provide a good model of concepts but I disagree that this is good reason to say that the notion of concept is scientifically idle.

It is better to say that, whereas an EYE-prototype, an EYE-exemplar and an EYE-theory are (empirically adequate) natural-kind terms relevant for supporting scientifically-interesting inductions about our capacity for categorising (what we take to be) members of the category of eyes, the concept EYE is a (metaphysically-motivated) natural-kind term that accounts for the truth maker of those inductions (i.e., inductions about categorisations are true, depending on whether or not an object categorised as eye is a token of the concept EYE). In a recent publication, Barrett (2015) introduces his discussion of concepts by pointing out that:

> There are concepts that all of us probably share at least in some way, such as the concepts PERSON, FOOD, MOTHER, WATER, and EYE. These concepts are not necessarily carbon-copy identical in everyone who possesses them; an ophthalmologist's concept EYE, for example, may differ substantially from that of a nonexpert. (p. 151)

In this passage, Barrett seems to overlook (or, perhaps, disregard) the distinction I have just made between the concept of X and what we know about the things that tokens of X represent. Bearing this distinction in mind, here is a possible reinterpretation of the quote above. First, what all of us probably share at least in some way is the very concept EYE (or PERSON, FOOD, MOTHER, WATER, etc.). EYE is an informationally-sustained HPC that supports inductive generalisations

[76] Some (e.g., Laurence and Margolis 1999) have gone on to suggest that a concept might have different kinds of structure (e.g., an atomic and various other kinds of structures) involved in the explanation of different psychological processes.

regarding the intentional explanations where that concept is involved. Second, what we do not share in terms of carbon-copy identical representations is what we know (or believe) about the things that tokens of EYE represent. Different interpreted computationally-sustained HPCs (e.g., Upper-MRs encoding information about prototypical eyes, particular eyes or causal-explanatory knowledge about eyes) can account for the concomitant systematic role of our background knowledge about the things in the extension of EYE. This background knowledge, as opposed to the concept EYE, is not the same in different contexts because, while we can all share the same concept, we never share exactly the same knowledge about the things that the tokens of that concept represent. Presumably, experts are in a better position to categorise things as eyes (e.g., the strange multiple eyes of some insects, perhaps even eyes of some creatures from outer space) because of the projectibility of their informed inductive generalisations about the members of the category of eyes. Finally, this does not mean that the ophthalmologist's concept EYE "differs substantially" from that of a nonexpert, because generalisations with respect to our capacities to categorise things as eyes and generalisations about the concept EYE are about different types of HPCs. In this sense, the fact the prototypes, exemplars and theories pick out HPCs that are relevant for categorisation does not challenge the scientific relevance of the term 'concept' because this term designates a different HPC in its own right.

3. Concluding remarks

The present thesis has aimed to challenge what is currently the most prominent eliminativist proposal regarding concepts and, as a result of that, defend the theoretical relevance of the term 'concept'. In order to accomplish this project, I have organised this work in three main stages.

In the first stage (Chapter 1), I explored different prominent eliminativist projects and showed that there are different ways of being an eliminativist regarding some type thing. In particular, eliminativists tend to advance arguments where a certain candidate for eliminativism (say, e.g., X, Y or Z) is said to fail to meet some

kind of conditions for theoretical or explanatory adequacy (say, e.g., X is theoretically inadequate due to the violation of certain metaphysical principle; Y is theoretically inadequate because it has no definite analysis; Z is theoretically inadequate because it is ontologically redundant). Accordingly, a general taxonomy was proposed, consisting of different categories and subcategories of eliminativist arguments and different categories of eliminativist claims (see Chapter 1, table 1). Relevant for the subsequent stages of the thesis, the proposed taxonomy allowed for the characterisation of a type of eliminativist projects that centres on the role of concepts that do not clearly designate a single class of things.

In the second stage (Chapters 2-4), I challenged the viability of what is currently the most prominent eliminativist proposal regarding concepts. In Chapter 2, I begin by providing an overview of the classical (or definitional) theory of concepts and the most prominent accounts of concepts that psychologists have developed as a reaction to that classical view. These alternative accounts identify concepts with one of three types of mental representations, namely, prototypes, exemplars and theories. Examining the main features and problems of each of these three types of theories allowed us to illustrate that their most serious challenges have a common origin, viz. the assumption that the way to individuate a concept is in virtue of its relations to other concepts. With this consideration in mind, Chapters 3 and 4 mount a direct attack on the type of concept eliminativism advocated by Machery. Thus, in Chapter 3, I take issue with Machery's general argument against the term 'concept', according to which cognitive scientists should abandon this term because it fails to pick out a natural kind (according to some conception of natural kinds) and keeping it would prevent scientific progress. Contrary to Machery, I argue both that this argument fails and that we should reject the elimination of concepts from the theoretical jargon of cognitive science. I offer two main reasons for this. On the one hand, Machery's theoretical framework for conceptual heterogeneity not only inherits many of the same problems facing the theories of concepts that Machery criticizes, but also introduces new problems that render his eliminativist proposal less viable. On the other, a reconstruction of a version of his argument bearing upon his so-called Heterogeneity Hypothesis allows to reveal an unwarranted assumption, namely, that an adequate theory of concepts must be committed to the view that concepts are some kind of *complex* mental representation. Chapter 4 goes on to contend that Machery's

alternative to concepts is significantly ill-equipped to solve the problem of intentional content. I argue that this is good reason to support the claim that eliminating the theoretical term ‘concept’, as opposed to keeping it, encumbers scientific progress.

In the final stage (Chapters 5-6), I advance a defence of the theoretical notion of concept by sketching an approach to natural kinds suitable for the current state of scientific immaturity of the study of cognition. Chapter 5 addresses the problem of natural kindness in philosophy of science with a focus on the study of concepts and attempts to characterise the sources of two apparent incompatible perspectives (or attitudes) towards natural kinds which I dub AUSTERE and INDULGENT, respectively. These perspectives are made manifest in the form of opposing attitudes with respect to a motley collection of background assumptions constraining the theorising about natural kinds (either explicitly or not). I argue that these attitudes amount to a set of unproductive tensions which demand some important reappraisal. This reappraisal should allow for a more productive approach to natural kinds regarding the study of concepts, namely one in which the advantages typically associated to one of the mentioned attitudes (i.e., AUSTERE or INDULGENT attitudes) are not viewed as a challenge for the other.

Consistent with these considerations, the present conclusion chapter has sketched a proposal regarding theoretical terms in cognitive science relevant for the task of identifying putative natural kinds and lawlike generalisations about them. According to this proposal, a suitable approach to natural kinds in a highly immature science of the mind must allow for the co-existence of different models of concepts and attitudes towards the study of concepts. I maintain that those otherwise opposing positions can productively fall under the qualified application of a single general conception of natural kinds, without disregarding their diverse motivations and explanatory interests. What I have specifically defended is a combined perspective on the notion of a concept as a natural-kind term, whereby the notion of Homeostatic Property Clusters can be heuristically applied both as per AUSTERE and INDULGENT in order to identify putative natural kinds designated by the term ‘concept’.

In this sense, the main contribution of the current work is that the view of scientifically relevant terms I favour not only undermines the premature elimination

of the notion of a concept, but it also offers common methodological space for issues raised by otherwise irreconcilable positions within philosophy of cognitive science, as those motivated by the several tensions characterised in terms of the AUSTERE and INDULGENT distinction.

Admittedly, the defence of the term ‘concept’ I have explored in this thesis is based on an approach to natural kindness (relevant for cognitive science) that is in need of further refinement. Future research may explore standards for the elimination of vague or imprecise theoretical terms in the context of the current state of development of cognitive science and, I believe, the proposed conception of natural kinds may subserve that purpose. For example, as I have pointed out, a vast amount of empirical evidence suggests that the notions of ‘prototype’, ‘exemplar’ and ‘theory’ designate putative psychological kinds that play some interesting role in certain cognitive processes. However, theorists have proposed different theories of each of those notions. Since there isn’t a clear consensus about which of the several theories of prototypes (or exemplars or theories) is to prevail over the others, Heterogeneity Eliminativism regarding the term ‘prototype’ (or ‘exemplar’ or ‘theory’) might apply. On pain of indiscriminate and unproductive eliminativism within a highly immature science of the mind, it is of paramount importance that we can figure out some minimum standards for the theoretical adequacy and elimination of core terms in cognitive science. Until those standards are worked out, it is safest to count on a theoretical vocabulary that not only supports further scientific progress in the study of concepts, but is also suited to resist the risks of hasty scientific elimination motivated by standards that may be irrelevant for a developing science of the mind. The goal of this thesis has been to show both that doing psychology without the term ‘concept’ is not an option and to propose a conception of theoretical terms that allows for further investigation into the role of concepts in our mental life.

Bibliography

Andreasen, R. (2004). The Cladistic Race Concept: A Defense, *Biology and Philosophy*, 19: 425–442.

Armstrong, S. L., Gleitman, L. R. and Gleitman, H. (1983). What Some Concepts Might Not Be, *Cognition* 13: 263–308.

Baker, L. R. (2007). *The Metaphysics of Everyday Life: An Essay in Practical Realism*, Cambridge: Cambridge University Press.

Barret, H. C. (2015). The evolution of conceptual design, in E. Margolis and S. Laurence (eds.), *The Conceptual Mind*, pp. 151–183, Cambridge, MA: The MIT Press.

Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts, in U. Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, pp. 101–140, Cambridge: Cambridge University.

Bateson, P. (1991). Are there Principles of Behavioural Development?, in P. Bateson (ed.), *The Development and Integration of Behaviour*, pp. 19–39. Cambridge: Cambridge University Press.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*, London: Routledge.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, London/Dordrecht: Kluwer.

Bird, A. and Tobin, E. (2015). Natural Kinds, in *The Stanford Encyclopedia of Philosophy* (Spring Edition), Retrieved May 16, 2016, from <http://plato.stanford.edu/archives/spr2016/entries/natural-kinds/>.

Blanchard, T. (2010). Default knowledge, time pressure, and the theory-theory of concepts, *Behavioral and Brain Sciences*, 33(2–3): 206–207.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds, *Philosophical Studies*, 61: 127–148.

Boyd, R. (1999). Kinds, complexity and multiple realization, *Philosophical Studies*, 95: 67–98.

Brigandt, I. (2009). Natural Kinds in Evolution and Systematics: Metaphysical and Epistemological Considerations, *Acta Biotheoretica*, 57(1): 77–97.

Brigandt, I. (2011). Natural kinds and concepts: a pragmatist and methodologically naturalistic account, in J. Knowles and H. Rydenfelt (eds.), *Pragmatism, Science and Naturalism*, pp. 171–196, Frankfurt am Main: Peter Lang Publishing.

Carey, S. (1985). *Conceptual Change in Childhood*, Cambridge: MIT Press.

Carey, S. (2009). *The origin of concepts*, Oxford: Oxford University Press.

Chakravartty, A. (2015). Scientific Realism, in *The Stanford Encyclopedia of Philosophy* (Fall Edition), Retrieved May 10, 2016, from <http://plato.stanford.edu/archives/fall2015/entries/scientific-realism/>.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes, *Journal of Philosophy*, 78 (2): 67–90.

Churchland, P. M. (1985). Conceptual progress and word/ world relations: in search of the essence of natural kinds, *Canadian Journal of Philosophy*, 15(1): 1–17.

Churchland, P. M. (1988). *Matter and Consciousness* (Revised Edition), Cambridge, MA: MIT Press.

Churchland, P. M. (1989). *A Neurocomputational Perspective*, Cambridge, MA: MIT Press.

Couchman, J. J., Boomer, J., Coutinho, M. V. C., and Smith, J. D. (2010). Carving nature at its joints using a knife called concepts, *Behavioral and Brain Sciences*, 33 (2–3), 207–208.

Cowie, F. (2009). Why Isn't Stich an ElimiNativist?, in D. Murphy and M. Bishop (eds.), *Stich and His Critics*, pp. 14–74, Oxford: Wiley-Blackwell.

Cracraft, J. (1983). Species Concepts and Speciation Analysis, in R. Johnston (ed.), *Current Ornithology*, pp. 159–187, New York: Plenum Press.

Cussins, A. (1990). The connectionist construction of concepts, in M. Boden (ed.), *The Philosophy of Artificial Intelligence*, pp. 368–440, Oxford: Oxford University Press.

Cussins, A. (1993). Nonconceptual Content and the Elimination of Misconceived Composites!, *Mind and Language*, 8 (2): 234–252.

Dawkins, R. (1982). *The Extended Phenotype*, Oxford: Oxford University Press.

De Sousa, R. (1984). The Natural Shiftiness of Natural Kinds, *Canadian Journal of Philosophy*, 14 (4), 561–580.

Dennett, D. (1987). *The Intentional Stance*, Cambridge, MA: MIT Press.

Douven, I. (2011). Abduction, in *The Stanford Encyclopedia of Philosophy*. Retrieved December 26, 2013, from <http://plato.stanford.edu/archives/spr2011/entries/abduction/>.

Dretske, F. (1981). *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

Dretske, F. (1988). *Explaining Behavior*, Cambridge, MA: MIT Press.

Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*, Cambridge MA: Harvard University Press.

Dupré, J. (2002). Is 'natural kind' a natural kind term?, *Monist*, 85 (1): 29–49.

Edwards, K. (2009). What concepts do, *Synthese*, 170: 289–310.

Edwards, K. (2010). Unity amidst heterogeneity in theories of concepts, *Behavioral and Brain Sciences*, 33(2–3): 210–211.

Elder, C. (2007). On the place of artifacts in ontology, in E. Margolis and S. Laurence (Eds.), *Creations of the Mind: Theories of Artifacts and Their Representation*, pp. 33–51, Oxford: Oxford University Press.

Fodor, J. (1974). Special Sciences or the disunity of the sciences as a working hypothesis, *Synthese*, 28: 97–115.

Fodor, J. (1981). The Mind-Body Problem, *Scientific American*, 244 (1):124—133.

Fodor, J. (1987). *Psychosemantics*, Cambridge: MIT Press.

Fodor, J. (1990). A Theory of Content, I & II, in his *A Theory of Content and Other Essays*, pp. 51–136, Cambridge, MA: MIT Press.

Fodor, J. (1994). Concepts: a potboiler, *Cognition*, 50 (1): 95–113.

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*, New York: Oxford University Press.

Fodor, J. (2003). 'Is it a Bird? Problems with Old and New Approaches to the Theory of Concepts', *Times Literary Supplement*, pp. 3–4.

Fodor, J. (2004). Having concepts: a brief refutation of the twentieth century, *Mind and Language*, 19 (1): 29–47.

Fodor, J. (2008). *LOT 2: The Language of Thought Revisited*, New York: Oxford University Press.

Fodor, J., Garrett, M., Walker, E., and Parker, C. (1980). Against Definitions, *Cognition*, 8(3): 263–367.

Fodor, J. and Lepore, E. (1992). *Holism: A Shopper's Guide*, Cambridge, MA: Basil Blackwell.

Fodor, J. and Lepore, E. (2002). *The Compositionality Papers*, Oxford: Oxford University Press.

Fodor, J. and Pylyshyn, Z. (2015). Minds Without Meanings: an Essay on the Content of Concepts, Cambridge, MA: The MIT Press.

Gelman, S. A., Coley, J. D. and Gottfried, G. M. (1994). Essentialist Beliefs in Children: The acquisition of Concepts and Theories, in L. A. Hirschfeld and S. A. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge: Cambridge University Press.

Gelman, S. A. and Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious, *Cognition*, 38(3), 213–244.

Gettier, E. L. (1963). Is Justified True Belief Knowledge?, *Analysis*, 23 (6): 121–123.

Gooding-Williams, R. (1998). Race, Multiculturalism and Democracy, *Constellations*, 5 (1): 18–41.

Gopnik, A. (1996). The Scientist as Child, *Philosophy of Science*, 63: 485–514.

Gopnik, A., and Meltzoff, A. (1997). *Words, Thoughts, and* Theories, Cambridge, MA: MIT Press.

Greenberg, M. and Harman, G. (2006). Conceptual Role Semantics, in E. LePore and B. Smith (eds.), *Oxford Handbook of Philosophy of Language*, pp. 295–322, Oxford: Oxford University Press.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66: 377–388.

Griffiths, P. E. (1997). *What Emotions Really Are*, Chicago: University of Chicago Press.

Griffiths, P. E. (2002). What is Innateness?, *The Monist* 85(1): 70–85.

Grush, R. (2001). The Philosophy of Cognitive Science. Retrieved May 16, 2016, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.405.3994&rep=rep1&type=pdf>.

Gunther, Y. H. (2003). General introduction, in Y. Gunther (ed.), *Essays on Nonconceptual Content*, pp. 1–19, Cambridge, MA: MIT Press.

Hacking, I. (1991). A tradition of natural kinds, *Philosophical Studies*, 61: 109–126.

Hacking, I. (1999). *The Social Construction of What?*, Cambridge, MA: Harvard University Press.

Hacking, I. (2007). Natural Kinds: Rosy Dawn, Scholastic Twilight, *Royal Institute of Philosophy Supplement*, 82 (61): 203–239.

Hampton, J. (1979). Polymorphous Concepts in Semantic Memory. *Journal of Verbal Learning and Verbal Behavior*, 18: 441–461.

Hampton, J. (1995). Testing the Prototype Theory of Concepts, *Journal of Memory and Language*, 34 (5): 686–708.

Hampton, J. (2006). Concepts as Prototypes, *Psychology of Learning and Motivation*, 46: 79–113.

Hampton, J. (2010). Concept talk cannot be avoided, *Behavioral and Brain Sciences*, 33 (2–3): 212–213.

Hanson, N. (1958). *Patterns of Discovery*, Cambridge: Cambridge University Press.

Harman, G. (1965). The Inference to the Best Explanation, *Philosophical Review*, 74: 88–95.

Harman, G. (1993). Meaning holism defended, in J. Fodor and E. Lepore (eds.), *Holism: a consumer update*, pp.163–171, Amsterdam: Rodopi.

Hayes, B. and Kearney, L. (2010). Defending the concept of "concepts," *Behavioral and Brain Sciences*, 33(2–3): 214–214.

Heider, E. (1972). Probabilities, Sampling, and Ethnographic Method: The Case of Dani Colour Names. Man, *The Journal of the Royal Anthropological Institute*, 7(3): 448–466.

Hendry, R. (2006). Elements, Compounds, and Other Chemical Kinds, *Philosophy of Science*, 73 (5): 864–875.

Hewson, C. (1994). Empirical evidence regarding the folk psychological concept of belief, in A. Ram and K. Eiselt (eds.), *Proceedings of the sixteenth annual conference of the Cognitive Science Society*, pp. 403–406, Hillsdale, NJ: Erlbaum.

Hill, C. (2010). I love Machery's book, but love concepts more, *Philosophical Studies*, 149 (3): 411–421.

Horgan, T. and J. Woodward. (1985). Folk Psychology is Here to Stay, *Philosophical Review*, 94: 197–225.

Hume, T., and Pazzani, M. (1995). Learning Sets of Related Concepts: A Shared Task Model, in *Proceedings of the Sixteen Annual Conference of the Cognitive Science Society*. Pittsburgh, PA: Lawrence Erlbaum. Retrieved May 16, 2016, from <http://www.ics.uci.edu/~pazzani/Publications/Hume-and-Pazzani-CogSci95.pdf>.

Jacobson, A. J. (2010). The faux, fake, forged, false, fabricated, and phony: Problems for the independence of similarity-based theories of concepts, *Behavioral and Brain Sciences*, 33(2–3): 215–215.

Keil, F. (1989). *Concepts, kinds, and cognitive development*, Cambridge, MA: MIT Press.

Keil, F., Levin, D., Gutheil, G. and Richman, B. (1999). Explanation, cause and mechanism: The case of contagion, in D. Medin and S. Atran (eds.), *Folkbiology*, pp. 285–320, Cambridge, MA: MIT Press.

Khalidi, M. (1998). Natural Kinds and Crosscutting Categories, *Journal of Philosophy*, 95: 33–50.

Kitcher, P. (1984). In Defense of Intentional Psychology, *Journal of Philosophy*, 81: 89–106.

Lahav, R. (1992). The Amazing Predictive Power of Folk Psychology, *Australasian Journal of Philosophy* 70: 99–105.

Laland, K. and Sterelny, K. (2006). Perspective: seven reasons (not) to neglect niche construction, *Evolution*, 60 (9): 1751–1762.

Lalumera, E. (2013). Concepts exist. More about Eliminativism, *Methode*, 2: 126–133.

LaPorte, J. (2004). *Natural Kinds and Conceptual Change*, Cambridge: Cambridge University Press.

Laurence, S. and Margolis, E. (1999). Concepts and cognitive science, in E. Margolis and S. Laurence (eds.), *Concepts: Core readings*, pp. 3–81. Cambridge, MA: MIT Press.

Machery, E. (2005). Concepts are not a natural kind, *Philosophy of Science* 72: 444–467.

Machery, E. (2009). *Doing Without Concepts*, New York: Oxford University Press.

Machery, E. (2010). Précis of *Doing Without Concepts*, *Behavioral and Brain Sciences*, 33 (2–3): 195–244.

Machery, E. (2015). By default, in E. Margolis and S. Laurence (eds.), *The Conceptual Mind*, pp. 567–588, Cambridge, MA: The MIT Press.

Machery, E., Mallon, R., Nichols, S. and Stich, S. P. (2004). Semantics, cross-cultural style, *Cognition*, 92 (3): B1–B12.

Mackie, J. (1977). *Ethics: Inventing Right and Wrong*, New York: Penguin.

Magnani, L. (2001). *Abduction, Reason, and Science. Processes of Discovery and Explanation*, New York: Kluwer Academic/Plenum Publishers.

Magnani, L. (2009). *Abductive Cognition: The Epistemological and Eco-cognitive Dimensions of Hypothetical Reasoning*. Berlin: Springer.

Magnus, P. (2012). *Scientific Enquiry and Natural Kinds: From Planets to Mallards*, Basingstoke: Palgrave Macmillan.

Mallon, R. (2006). Race: Normative, Not Metaphysical or Semantic, *Ethics* 116 (3): 525–551.

Malt, B. (1994). Water Is Not H2O, *Cognitive Psychology*, 27 (1): 41–70.

Malt, B. and Johnson, E. (1992). Do Artifact Concepts Have Cores?, *Journal of Memory and Language*, 31 (2): 195–217.

Mameli, M. and Bateson, P. (2005). Innateness and the Sciences, *Biology and Philosophy* 21 (2): 155–88.

Margolis, E. (1998). How to Acquire a Concept, *Mind and Language*, 13 (3): 347–369.

Margolis, E. and Laurence, S. (2007). The Ontology of Concepts — Abstract Objects or Mental Representations?, *Noûs*, 41 (4): 561–93.

Margolis, E. and Laurence, S. (2010). Concepts and theoretical unification, *Behavioral and Brain Sciences*, 33(2–3): 219–220.

Margolis, E. and Laurence, S. (2011). Learning Matters: The Role of Learning in Concept Acquisition, *Mind and Language*, 26 (5): 507–539.

Matsuka, T. and Sakamoto, Y. (2007). A Cognitive Model That Describes the Influence of Prior Knowledge on Concept Learning, in *Artificial Neural Networks, Lecture Notes on Computer Science (LNCS)* 4668: 912–921, Berlin: Springer-Verlag.

Mayr, E. (1969). *Principles of Systematic Zoology*, New York: McGraw-Hill.

McGrath, M. (2005). "No Objects, no Problem?," *Australasian Journal of Philosophy*, 83: 457–486.

Medin, D. (1989). Concepts and Conceptual Structure, *American Psychologist*, 44 (12): 1469–1481.

Medin, D. and Ortony, A. (1989). Psychological Essentialism., in S. Vosniadou and A. Ortony (Eds.), *Similarity and Analogical* Reasoning, pp. 179–195. New York: Cambridge University Press.

Medin, D., and Schaffer, M. (1978). A Context Theory of Classification Learning. *Psychological Review*, 85: 207–238.

Medin, D. and Schwanenflugel, P. (1981). Linear separability in classification learning, *Journal of Experimental Psychology: Human Learning and Memory*, 7 (5): 355–368.

Miller, G. (2003). The cognitive revolution: a historical perspective, Trends in Cognitive Sciences, 7 (3): 141–144.

Millikan, R. (1989). Biosemantics, *Journal of Philosophy*, 86: 281–97.

Millikan, R. (1998). A Common Structure for Concepts of Individuals, Stuffs, and Real Kinds: More Mama, More Milk, and More Mouse, *Behavioral and Brain Sciences*, 21: 55–65.

Murphy, G. (2002). *The Big Book of Concepts*, Cambridge, MA: MIT Press.

Murphy, G. and Medin, D. L. (1985). The role of theories in conceptual coherence, *Psychological Review* 92: 289–316.

Needham, P. (2000). What is Water?, *Analysis*, 60: 13–21.

Nersessian, N. (1999). Model-based reasoning in conceptual change, in Magnani, L., Nersessian, N., and Thagard, P. (eds.), *Model-Based Reasoning in Scientific Discovery*, pp. 5–22, New York: Kluwer Academic/Plenum Publishers.

Nersessian, N. (2008). *Creating Scientific Concepts*, Cambridge, MA: MIT Press.

Nisbett, R., Peng, K., Choi, I. and Norenzayan, A. (2001). Culture and Systems of Thought: Holistic Versus Analytic Cognition, *Psychological Review*, 108 (2): 291–310.

Papineau, D. (2015). Naturalism, in *The Stanford Encyclopedia of Philosophy* (Fall Edition), Retrieved May 10, 2016, from <http://plato.stanford.edu/entries/naturalism/>.

Pazzani, M. (1991). Influence of Prior Knowledge on Concept Acquisition: Experimental and Computational Results, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17 (3): 416–432.

Peacocke, C. (1992). *A study of Concepts*, Cambridge, MA: MIT Press.

Peacocke, C. (1996). Can possession conditions individuate concepts?, *Philosophy and Phenomenological Research*, 56: 433–60.

Peacocke, C. (2009). Concepts and Possession Conditions, in B. McLaughlin, A. Beckermann, and S. Walter (eds.), *The Oxford Handbook of Philosophy of Mind***,** pp. 437–456. Oxford: Oxford University Press.

Piccinini, G. (2015). Computation in Physical Systems, in *The Stanford Encyclopedia of Philosophy* (Summer Edition), Retrieved April 26, 2016, from <http://plato.stanford.edu/archives/sum2015/entries/computation-physicalsystems/>.

Piccinini, G. and Scarantino, A. (2010). Computation vs. information processing: why their difference matters to cognitive science, *Studies in History and Philosophy of Science*, 41(3), 237–246.

Piccinini, G. and Scott, S. (2006). Splitting concepts, *Philosophy of Science*, 73 (4): 390–409.

Posner, M., and Keele, S. (1968). On the Genesis of Abstract Ideas, *Journal of Experimental Psychology*, 77: 353–363.

Pöyhönen, S. (2013). *Natural Kinds and Concept Eliminativism*, in V. Karakostas and D. Dieks (eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings*, 2: 167–179, Cham, Switzerland: Springer International Publishing.

Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*, Cambridge, MA: MIT Press.

Prinz, J. (2005). The return of concept empiricism, in H. Cohen and C. Lefebvre (eds.), *Handbook of categorization in cognitive science*, pp. 679–695, Oxford: Elsevier.

Prinz, J. (2011). Has Mentalese earned its keep? On Jerry Fodor's LOT 2, *Mind*, 120(478): 485–501.

Prinz, J., and Clark, A. (2004). Putting Concepts to work: some thoughts for the twenty first century, *Mind and Language*, 19 (1): 57–69.

Putnam, H. (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, H. (1975). The Meaning of "Meaning," in H. Putnam, *Philosophical Papers*, vol. 2: *Mind, Language, and Reality*, pp. 215–271, Cambridge: Cambridge University Press.

Pylyshyn, Z. (2002). Mental imagery: In search of a theory, *Behavioral and Brain Sciences*, 25 (2): 157–182.

Quine, W.V.O. (1951/1980). Two Dogmas of Empiricism. In *From a Logical Point of View*, pp. 20–46,. Cambridge, MA: Harvard University Press.

Quine, W.V.O. (1969). Natural Kinds, in *Ontological Relativity and Other Essays*, New York: Columbia University Press.

Ramsey, W. (2013). Eliminative Materialism, in *The Stanford Encyclopedia of Philosophy* (Summer Edition), Retrieved March 16, 2015, from <http://plato.stanford.edu/archives/sum2013/entries/materialism-eliminative/>.

Rey, G. (1983). A reason for doubting the existence of consciousness, in R. Davidson, G. Schwartz, and D. Shapiro (eds.), *Consciousness and Self-regulation,* Vol. 3. New York, Plenum.

Rey, G. (1985). Concepts and conceptions: A reply to Smith, Medin and Rips, *Cognition*, 19(3): 297–303.

Rey, G. (1998). Eliminativism, in *Routledge Encyclopedia of Philosophy*. Retrieved May 16, 2016, from <https://www.rep.routledge.com/articles/eliminativism/v-1/>.

Rey, G. (2009). Review of E. Machery, *Doing without Concepts*, in *Notre Dame Philosophical Reviews*. Retrieved July 24, 2013 from < http://ndpr.nd.edu/news/24087-doing-without-concepts/>

Rey, G. (2010). Concepts versus conceptions (again), *Behavioral and Brain Sciences*, 33(2–3): 221–222.

Rips, L. (1995). The current status of the research on concept combination, *Mind and Language*, 10: 72–104.

Rosch, E. (1973). Natural categories, *Cognitive Psychology*, 4 (3): 328–350.

Rosch, E. (1978). Principles of Categorization, in E. Rosch and B. Lloyd (eds.), *Cognition and Categorization*, pp. 27–48, Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosch, E., and Mervis, C. B. (1975). Family Resemblances: Studies in the Internal Structure of Categories, *Cognitive Psychology*, 7: 573–605.

Samuels, R. (2009). Delusions as a natural kind, in M.R. Broome and L. Bortolotti (eds.), *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, pp. 49–82, Oxford: Oxford University Press.

Samuels, R. and Ferreira, M. (2010). Why don't concepts constitute a natural kind?, *Behavioral and Brain Sciences*, 33(2–3): 222–223.

Scanlon, T. (2014). *Being Realistic About Reasons*, Oxford: Oxford University Press.

Scarantino, A. (2010). Evidence of coordination as a cure for concept eliminativism, *Behavioral and Brain Sciences*, 33(2–3): 223–224.

Schneider, S. (2010). Conceptual atomism rethought, *Behavioral and Brain Sciences*, 33(2–3): 224–225.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*, *Cambridge*, MA: MIT Press.

Searle, J. (1983). *Intentionality: an Essay in the Philosophy of Mind*, Cambridge: Cambridge University Press, 1983.

Sesardic, N. (2010). Race: A Social Destruction of a Biological Concept, *Biology and Philosophy*, 25: 143–162.

Shafer-Landau, R. (2003). *Moral Realism: A Defense*, Oxford: Oxford University Press.

Simon, H. and C. Kaplan (1989). Foundations of cognitive science, in M.I. Posner (ed.), *Foundations of Cognitive Science*, pp. 1–47, Cambridge MA: MIT Press.

Sloman, S. (1993). Feature-Based Induction, *Cognitive Psychology*, 25 (2): 231–280.

Smart, J. (1959). Sensations and Brain Processes, *Philosophical Review*, 68: 141–156.

Smith, E., and Medin, D. L. (1981). *Categories and concepts*, Cambridge, MA: Harvard University Press.

Smith, E. and Sloman, S. (1994). Similarity- versus rule-based categorization, *Memory and Cognition*, 22: 377–386.

Spalding, T. L. and Murphy, G. L. (1999). What Is Learned in Knowledge-Related Categories? Evidence from Typicality and Feature Frequency Judgments, *Memory and Cognition*, 27 (5): 856–867.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*, Cambridge MA: MIT Press.

Stich, S. (1996). *Deconstructing the Mind*, New York: Oxford University Press.

Strohminger, N. and Moore, B. W. (2010), Banishing the thought, *Behavioral and Brain Sciences*, 33 (2–3), 225–226.

Thagard, P., Magnani, L. and Nersessian, N. (1999). *Model-based Reasoning in Scientific Discovery*, New York/London: Kluwer Academic/Plenum Publishers.

The top 10 film moments. (2000). Retrieved February 5, 2014, from <http://www.theguardian.com/film/2000/feb/06/top-10-film-moments-usual-suspects-psycho>.

Tversky, A. (1977). Features of similarity, *Psychological Review*, 84 (4): 327–352.

Unger, P. (1979). There are no ordinary things. *Synthese*, 41: 117–54.

Van Fraassen, B. (1980). *The Scientific Image*, Oxford University Press.

Van Inwagen, P. (1990). Artifacts, in his *Material Beings*, pp. 124–130. Ithaca: Cornell University Press.

Virtel, J. and Piccinini, G. (2010). Are prototypes and exemplars used in distinct cognitive processes?, *Behavioral and Brain Sciences*, 33(2–3): 226–227.

Weisberg, M., Needham, P. and Hendry, R. (2011). Philosophy of Chemistry, in *The Stanford Encyclopedia of Philosophy* (Winter Edition). Retrieved May 16, 2016, from <http://plato.stanford.edu/archives/win2011/entries/chemistry/>.

Weiskopf, D. (2009). The plurality of concepts, *Synthese*, 169: 145–173.

Weiskopf, D. (2010). The theoretical indispensability of concepts, *Behavioral and Brain Sciences*, 33(2–3): 228–229.

Wittgenstein, L. (1953/1958). *Philosophical Investigations*, 3rd edition, G.E.M. Anscombe (trans.), Oxford: Blackwell.

Zack, N. (2002). *Philosophy of Science and Race*, New York: Routledge.