THE UNIVERSITY
of EDINBURGH

FROM FOLK PSYCHOLOGY TO COGNITIVE ONTOLOGY

JOE DEWHURST

PhD in Philosophy

The University of Edinburgh

2017

Abstract

This thesis examines the relationship between folk psychology and scientific psychology, and argues that the conceptual taxonomy provided by the former is unsuitable for fine-grained cognitive scientific research. I avoid traditional eliminativism by reserving a role for folk psychology as a socio-normative discourse, where folk psychological concepts primarily refer to behaviour rather than to mental states, and also exert a regulative influence on behaviour.

In the first half of this thesis I develop a positive account of folk psychology as a broad discourse that includes mental state attributions, behavioural predictions, narrative competency, and regulative mechanisms. In the second half I argue that the conceptual taxonomy provided by this discourse has led to theoretical confusions in both philosophy and cognitive science, and I propose a systematic methodology for developing a novel 'cognitive ontology' that is better suited for contemporary scientific research.

What is folk psychology? In **chapter 1** I survey the history of the term folk psychology and demonstrate that the term only really came into general usage following the work of Fodor and Churchland in the 1970s and 80s. I also argue that it is a mistake, stemming from this era, to identify folk psychology exclusively with propositional attitude psychology, which is just one particular way in which the folk might understand one another.

If folk psychology is not just propositional attitude psychology, what else might it be? In **chapter 2** I consider what I call the 'universality assumption', i.e. the assumption that folk psychological intuitions are shared across all cultures and languages. If this assumption were justified then it might provide partial support for the claim that folk psychology presents an accurate account of human cognition. However, there is significant evidence of variation in folk psychological intuitions, suggesting that folk psychology might be at least partially biased by cultural and linguistic influences.

If folk psychology is not the same in every culture, how come it is so successful at predicting behaviour? In **chapter 3** I look at various ways in which folk psychological discourse can play a regulative or normative role by exerting an influence on our behaviour. This role helps to explain how folk psychology can be predictively successful even if it fails to accurately describe the fine-grained details of human cognition, as via regulative mechanisms it is able to become a kind of self-fulfilling prophecy.

How well does folk psychology match up with our scientific understanding of cognition? In **chapter 4** I present evidence of cases where folk psychological concepts have served to mislead or confuse theoretical debates in philosophy of mind and cognitive science. I consider several case studies, including the false belief task in social cognition, the taxonomisation of sensory modalities, the extended cognition debate, and the recently emerging 'Bayesian brain' hypothesis.

If folk psychological concepts do not refer to entities in our scientific theories, then what do they refer to? In **chapter 5** I examine the status of folk psychological kinds as natural kinds, and argue that even under a very liberal account folk psychological kinds probably do not constitute viable scientific kinds. However, due to the regulative mechanisms described in chapter 3, they do constitute what Hacking has described as 'human' or 'interactive' kinds, which exhibit complex looping effects.

What kinds of concepts should cognitive science use, if not folk psychological concepts? Finally, in **chapter 6** I look at recent developments in 'cognitive ontology' revision and argue that we should adopt a systematic methodology for constructing novel concepts that better reflect our current best understanding of cognitive systems. In closing I consider the relationship between these novel concepts and the ontology presented by folk psychological discourse.

Acknowledgements

My parents Mark and Lesley have supported me throughout my time as a student, and I thank them for always encouraging me and my siblings to pursue our interests, no matter how unlikely the career prospects.

I am grateful to Guy Goodwin and Tim Lawrence for first introducing me to philosophy and psychology, respectively, and for instilling in me a light-hearted but nonetheless rigorous approach to my studies.

My work on this thesis was supported by a scholarship from the Carnegie Trust for the Universities of Scotland. The title is a shameless adaption of Stich's *From Folk Psychology to Cognitive Science*, which has heavily influenced my work.

Dave Ward and Suilin Lavelle have excelled, both as supervisors and as mentors, and I hope one day to pass on all they have given me to students of my own. Suilin also supervised my undergraduate dissertation, and taught me all that I know about social cognition. Without her influence I may never have considered postgraduate study.

Rosa Hardt has been a steady companion and friend over the past four years, as well as an antidote to my reductionist tendencies. The other MLECers (and honorary MLECers) have also influenced me greatly, and I hope we stay in touch for many years to come.

Ana-Maria Cretu spent many hours discussing natural kinds with me, and Jonny Lee and Dan Calder kept me sane with philosophical conversations about something other than folk psychology. Many people have read drafts of these chapters for me, but I would especially like to thank Marco Viola for his comments and suggestions. Fiona Doherty and her tomatoes gave me the final push I needed to get this thesis finished.

Emma persuaded me to work on this thesis, and without her I would not be where I am today. I am eternally grateful to her for putting up with me for so long.

This thesis is dedicated to the memory of my grandmother Brenda Croydon, who passed away on March 1st, 2015. She was a keenly intelligent and compassionate woman, and an inspiration to us all.

Declaration

This thesis has been composed by me and is entirely my own work. It has not been submitted for any other degree or professional qualification. Much of the material in section 4.4 is forthcoming as "Folk Psychology and the Bayesian Brain", in Metzinger & Wiese (eds.), *Philosophy and Predictive Processing*.

Joe Dewhurst
3rd February 2017

"So in the end, when one is doing philosophy, one gets to the point where one would like just to emit an inarticulate sound."

Ludwig Wittgenstein

Introduction: Two Discourses, Both Alike in Dignity

This thesis is concerned with the interaction between two ways of talking about human behaviour, the everyday and the scientific. I will refer to these two ways of talking as folk psychological discourse and cognitive scientific discourse respectively. The diagram below gives a schematic overview of their interactions, which will be described in more detail in the following chapters. The remainder of this section provides some initial definitions to help orientate the reader, and an overview of the chapters that follow.

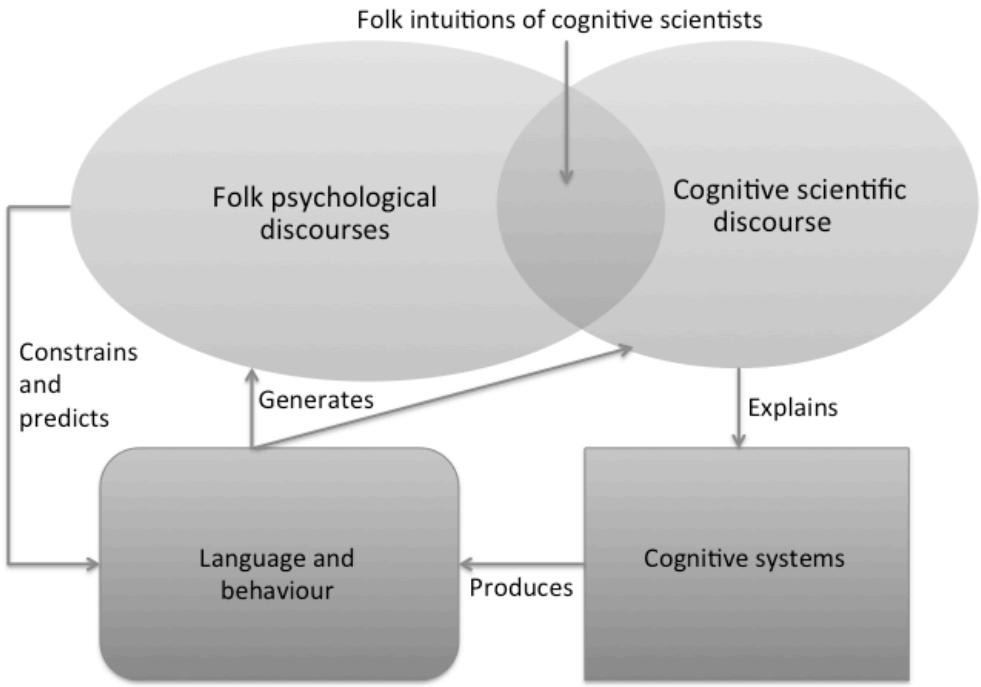

Figure 0.1. The relationship between folk psychological and cognitive scientific discourse.

'Folk psychological discourse' describes the everyday and intuitive ways that we talk about the behaviour of ourselves and of other people (and sometimes animals and other non-people such as machines or natural phenomena). This includes not only the attribution of beliefs, desires, and other propositional attitudes, but also behavioural predictions, narratives and other alternative ways of describing and

explaining behaviour. These discourses at least partially overlap with cognitive scientific discourse, especially when it comes to the scientific investigation of social behaviour, but also when scientists' own folk intuitions exert an influence on their research. It is the scope and impact of this influence that I am primarily concerned with in the second half of this thesis.

'Cognitive scientific discourse' describes the language and terminology used by scientific researchers to explain the behaviour of cognitive systems (of which humans are just one especially interesting example). This covers a wide range of disciplines, spanning from those that are concerned with the fine-grained details of the brain and nervous system, such as cognitive neuroscience, to those that are concerned with coarse-grained characterisations of personal level behaviour, such as social psychology. As such, cognitive scientific discourse encompasses multiple, potentially incompatible characterisations of how cognitive systems work. In addition, some of these characterisations will be compatible with folk psychological discourse, whilst other might not be. Cognitive scientific discourse can also influence folk psychological discourse (e.g. the popularisation of psychoanalytic theories of the unconscious), although influence in this direction will not be the primary focus of this thesis.

'Cognitive systems' refers to any system studied by cognitive science (this description is transparently circular, but it is intended to be descriptive rather than carrying any metaphysical importance). This includes the human brain and nervous system, bodily behaviour produced by the brain and nervous system (including linguistic utterances), and interactions between such behaviour and the external world (which can include other cognitive systems). Cognitive systems are (ideally) explained by cognitive scientific discourse, and produce language and behaviour (or are perhaps part constituted by them; see below).

'Language and behaviour' refers to the outputs generated by cognitive systems. There is, of course, a high degree of overlap between a cognitive system and the behaviour that it produces, and cognitive science studies both of these phenomena, but for the purposes of this orientation it is useful to introduce an artificial distinction between the two. Language and behaviour are constrained and predicted by folk psychological discourse, produced by cognitive systems, and generate, amongst other things, both folk psychological discourse and cognitive scientific discourse. This introduces recursive, looping effects that make any study of these systems especially complicated.

Each chapter of this thesis will examine one part of the above diagram in more detail. The thesis as a whole is split into two halves. In the first half (chapters 1, 2, and 3) I will present a positive account of folk psychology as a complex socio-cultural discourse, one that is capable of predicting and explaining behaviour, and that also exerts a regulative influence. In the second half (chapters 4, 5, and 6) I will argue that despite the many successes of folk psychology, it is often ill-suited for usage in philosophy and cognitive science, and propose a new methodology for developing novel, non-folk psychological concepts for the study of cognition. A chapter-by-chapter breakdown can be found below.

Chapter 1 will focus on explicating exactly what is meant by folk psychological discourse, and clarifying several distinct uses of the term folk psychology in contemporary philosophy. **Chapter 2** will discuss cross-cultural variation in folk psychological discourse, and challenge the assumption that folk psychological intuitions are universal. **Chapter 3** will look at the relationship between folk psychological discourse and language and behaviour, and argue that folk psychology should be understood as having a regulative as well as a predictive role with regard to behaviour. By the end of the first half I hope to have clarified what I think folk psychology is, and how I think it can contribute to predictions and explanations of our behaviour. Between the two halves there is a short interlude that

summarises my account of folk psychology, and compares it to several similar accounts.

The second half opens with **Chapter 4**, which will explain what is meant by cognitive scientific discourse, and then explore a number of case studies where folk psychological intuitions and terminology appear to have confounded theoretical issues in philosophy and cognitive science. **Chapter 5** will examine the status of folk psychological kinds as natural kinds, and consider the role of natural kinds in cognitive science. **Chapter 6** will look at recent work on cognitive ontology revision, and propose that we adopt a mechanistic approach to cognitive ontology, wherein folk psychological concepts and explanations can sometimes serve as sketches of mechanisms. At the end of the second half I will conclude by considering in more detail the relationship between folk psychological and cognitive scientific discourse.

The aim of this thesis is twofold. Firstly, to further our understanding of 'folk psychology', and to clarify the ways in which that term is currently used in philosophy and cognitive science. Secondly, to ask whether folk psychological concepts, i.e. intuitive ways of categorising mind and behaviour, are really suitable for technical application in philosophy and cognitive science. The upshot of this latter question will be that often they are not, and that therefore we should investigate ways of developing novel conceptual taxonomies that better reflect our growing understanding of how cognitive systems function. The upshot of the first aim, however, will be that despite this conceptual mismatch between folk psychology and cognitive science, there remains an important role for folk psychology as a personal level, social and normative discourse that is valuable in its own right. Thus whilst at times I might seem to endorse a form of eliminativism with regard to folk psychology, I only advocate the elimination of folk psychological concepts *from scientific discourse* (and even then, only when it is actually the case that such concepts are being misapplied).

My hope is that this thesis will be both useful and interesting for anyone working in the cognitive sciences, broadly construed. The first half is perhaps of

especial interest to those working in social cognition, as I argue against conflating the content of folk psychological discourse with the structure of social cognitive mechanisms, and propose a distinctive and novel way of characterising folk psychology. The second half is likely to be of more general interest, as it deals with a number of cases where folk psychological intuitions and concepts might be being misapplied, or might otherwise turn out to be unhelpful. Finally, by suggesting that folk psychological explanations might sometimes serve as sketches of mechanisms, I hope to contribute to the development of mechanistic explanations in cognitive science.

Chapter 1 – Folk Psychological Discourse

Talk of folk psychology is ubiquitous in contemporary philosophy, and yet despite this ubiquity it is not always clear what it is meant to refer to. Is it a theory of how the mind works, a non-theoretical body of knowledge, or some other kind of mechanism or process? The aim of this chapter is to clarify what we mean when we talk about folk psychology, and to give an initial description of several distinct phenomena that can all be referred to as folk psychological. I will also provide an overview of the historical usage of the term 'folk psychology', both in philosophy and in cognitive science, and describe in more detail the amalgamation of folk knowledge and social practices that I will be referring to as folk psychological discourse. Subsequent chapters will consider the impact that this discourse has had on research in philosophy and cognitive science, and what (if anything) we should be doing to counteract this impact.

I take it that folk psychology, in its most general sense, simply refers to "whatever interpersonal understanding consists of" (Ratcliffe 2009: 380), but there are at least two ways of characterising what this might be. 'Folk psychology' is used interchangeably in current literature to refer to two distinct phenomena: the explicit way that we talk about and understand the behaviour and cognition of other people, which I call **folk psychological discourse** and Ratcliffe calls "folk folk psychology" (*ibid*: 381), and the implicit processes that facilitate social interaction, which I call **social cognitive mechanisms**. Stich & Ravenscroft (1994) make a similar distinction between 'internal' and 'external' folk psychology, where internal folk psychology refers to the "tacit rules and generalizations [that] play a central role in explaining folk psychological capacities" (*ibid*: 459), whilst external folk psychology refers to the "consciously accessible consequences" (*ibid*) of these rules and generalizations. Ignoring, for the moment, the question of whether social cognitive mechanisms are in fact best characterised as a set of tacit rules and generalizations, let me try and explicate this distinction in some more detail.

Whenever I interact with another person, I am able to more-or-less effortlessly predict subtle changes in their behaviour, and subsequently respond in the right sort of way. For example, I might step out of the way as someone walks towards me, or stop talking when I sense that they are about to say something. Of course, these predictions are often unsuccessful; I might bump into someone on the street, or talk over someone when they try to say something; but when these predictions are at their most successful I am almost unaware that I am making them, and no conscious effort goes into the process of working out what someone else is about to do. This capacity (or set of capacities) for effortlessly predicting and responding to the behaviour of other people is underpinned by what I 'social cognitive mechanisms', what Stich & Ravenscroft call 'internal folk psychology', and what is more typically (and confusingly) referred to simply as 'folk psychology' or 'theory of mind'.

At the same time, whenever I interact with another person, I am able to consciously reflect on their behaviour, and if prompted could offer an explanation of why they behaved like they did. I can also consciously reflect on my own behaviour. For example, I might say that I bumped into that stranger because we were both in a hurry, or that you and I spoke over one another because we were both so excited about the conversation that we were having. These explanations may or may not be true, but regardless they seem to serve an important social function, and are philosophically and psychologically interesting in their own right. It is this explicit capacity to reflect on and explain behaviour that I call 'folk psychological discourse', Stich & Ravenscroft call 'external folk psychology', and Ratcliffe calls 'folk folk psychology'. More typically it is *also* referred to as 'folk psychology', bundling together the two phenomena that I have described here, which I think is the source of many confusions and misunderstandings. Thus I will make a concerted effort to keep the two distinct, using the terms *social cognitive mechanisms* and *folk psychological discourse*, or just *social cognition* and *folk psychology*, unless otherwise specified.

These two phenomena may overlap to a greater or lesser extent, and each can be further subdivided, but they must be kept at least in principle distinct from one

another. This is essential if we are going to be able to ask questions such as whether folk psychological discourse and/or social cognitive mechanisms are culturally universal, without the assumption that the answer will be the same in both cases. It is also essential if we want to question the scientific accuracy of folk psychological discourse, without denying that social cognitive mechanisms exist, or consider whether folk psychological discourse might serve some other, non-scientific role. The rest of this chapter will explore the distinction between folk psychological discourse and social cognition in more detail, starting with a historical overview of the term folk psychology in philosophy and cognitive science before proceeding to explain how I think the distinction between social cognition and folk psychological discourse should be understood.

1.1 – History of Folk Psychology

Despite its ubiquity, the term folk psychology is relatively young, making it possible to present a more-or-less complete history of its usage, in both of the senses described previously. This will help clarify the distinctions that I introduced above, and which I explain in more detail in the rest of this chapter. My history of the term folk psychology is divided into six eras, some of which overlap chronologically, but each of which corresponds to a particular way in which the term has been understood or applied. The six eras are as follows:

1. *Völkerpsychologie* and Propositional Attitudes (Early 20th century)
2. The Myth of Jones (1949-1963)
3. Folk Theory and Eliminativism (1966-1981)
4. Classical Cognitive Science (1968-1987)
5. Social Cognition (1978-present day)
6. Connectionism and Embodiment (1986-present day)

It is worth noting that the term folk psychology is only used in its modern sense from about 1980 onwards. Prior to this the term is either used to refer to a distinct

methodology, as in the work of Wilhelm Wundt, or else it is replaced with the term common-sense psychology, which I am assuming has roughly the same meaning. Thus a history of the term folk psychology is also a history of the term common-sense psychology.

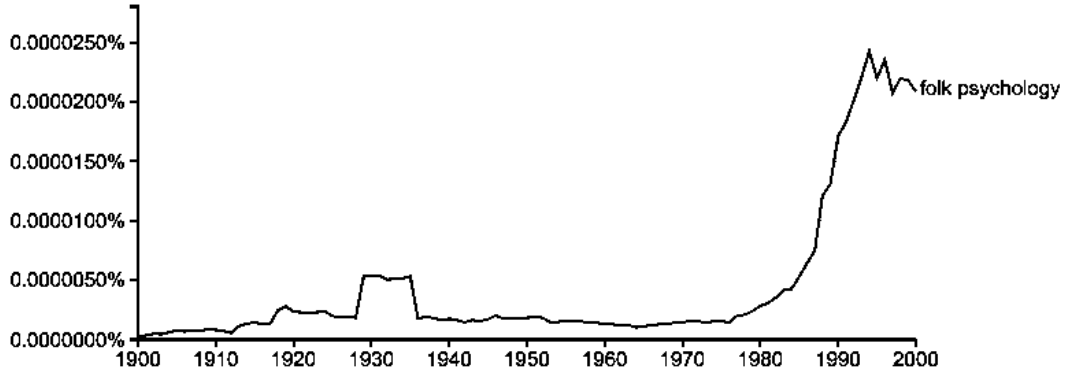

Figure 1.1. Google Ngram for "folk psychology" from 1900-2000.

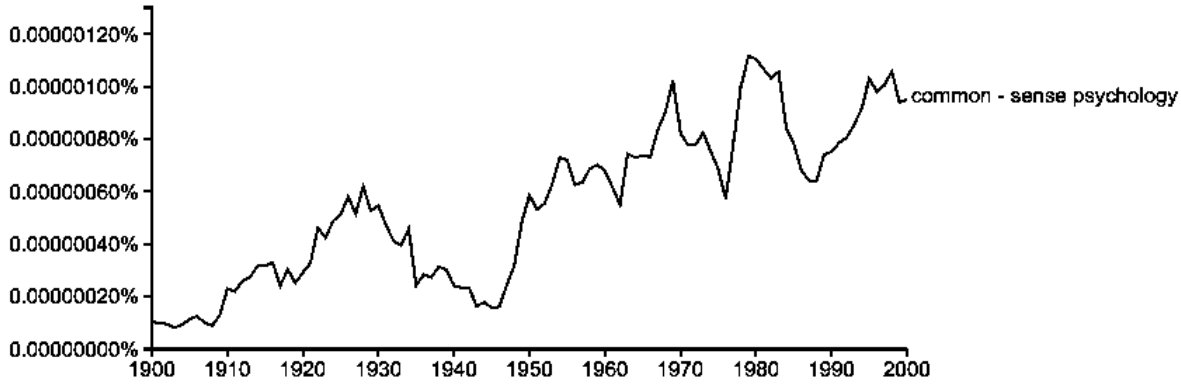

Figure 1.2. Google Ngram for "common-sense psychology" from 1900-2000.

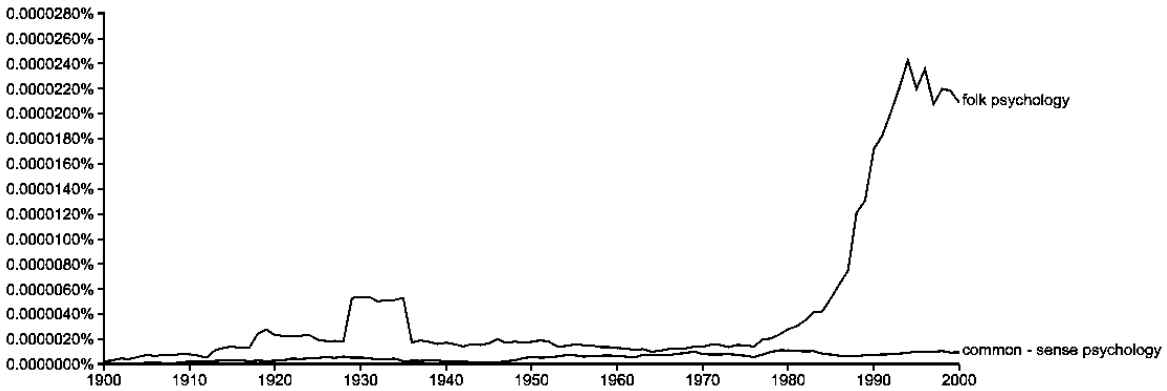

Figure 1.3. Google Ngram for both terms from 1900-2000.

The diagrams above were created using Google Ngram to show the frequency of the phrases "folk psychology" and "common-sense psychology" in English language books published during the 20th century.[1] Note the spike in "folk psychology" beginning around 1980 (fig. 1.1), which coincides with the Churchlands' work on eliminative materialism (see section 1.1.3). The earlier blip from 1928-36 is probably the result of translations of the German word *Völkerpsychologie* (see section 1.1.1). The graph for "common-sense psychology" (fig. 1.2) is messier but shows a similar spike starting around 1949, when Gilbert Ryle's *The Concept of Mind* was published (see section 1.1.2). The graph for both (fig. 1.3) is included to indicate the relative frequencies of the two terms. These diagrams are merely suggestive, but the patterns they reveal seem to correlate well with the eras that I have chosen to focus on.

1.1.1 – *Völkerpsychologie* and Propositional Attitudes (Early 20th century)

The term folk psychology first appears as a translation of the German word *Völkerpsychologie,* used by the early experimental psychologist Wilhelm Wundt to describe the kind of social and cultural enquiry that he thought should accompany psychological experimentation (Wundt 1912/1916; Kim 2008: sec. 6). This would involve a wide-ranging synthesis of techniques and data from disciplines such as history, linguistics, and anthropology, but only insofar as they cast light on the psychological processes that he was primarily interested in. We can find echoes of it in modern psychological anthropology, which is "the study of the behaviour, experience and development of individuals in relation to the institutions and ideologies of their sociocultural environments" (LeVine 2010: 1). Later in this chapter (1.3) I will argue that folk psychology is best characterised as a complex discourse, and one consequence of this characterisation is that something like the anthropological technique suggested by Wundt's *Völkerpsychologie* is in fact required if we are to fully understood how the 'folk' think about other minds.

[1] This use of Google Ngram was inspired by Andow (2015), and facilitated by Michel *et al* (2011).

Chapter 2 will pursue this methodology in more detail, and consider recent evidence of cross-cultural variation in folk psychological discourse.

Wundt's work on scientific psychology proper also indicates an awareness of some of the issues associated with drawing on what we now refer to as folk psychology in an experimental context. In an intriguing passage that I quote in its entirety below, Danziger seems to suggest that Wundt, the father of modern psychology, was steadfastly opposed to any use of folk psychological terminology or concepts in the scientific context.

> [In the introduction to his *Principles of Physiological Psychology,*] Wundt notes that ordinary language provides us with certain terms for classifying psychological events, e.g. feeling, understanding, sensibility and also memory, which, in pre-scientific psychology, are taken to identify distinct faculties or mental powers. Such ordinary-language psychological categories are dangerous for the project of a scientific psychology because they tend to confound descriptions and explanation. Scientific psychology has to make a clear separation between categories of observed phenomena to be explained and theories that do the explaining. (Danziger 2008: 126)

Danziger is primarily concerned with Wundt's influence on the early days of the scientific study of memory. Immediately after the passage quoted above, he indicates that it was precisely this dismissal of folk concepts that prevented Wundt from engaging in any serious way with the study of memory. According to Wundt, memory was only a "surface product generated by more fundamental psychological processes" (Danziger 2008: 126), and as such did not qualify as a phenomenon worthy of detailed scientific study. This attitude echoes that of the later eliminative materialists, whom I discuss in more detail in 1.1.3. So even here, in the pre-history of both folk and scientific psychology, there was a close but not always friendly relationship between the two.

Wundt's word *Völkerpsychologie* appears to be relatively unconnected to our modern usage use of the term folk psychology (Ratcliffe 2007: 42-3), but it nonetheless provides an interesting counterpoint to the received view of folk

psychology as propositional attitude psychology. This received view dates back to Frege and Russell's work on propositional attitude reports, which introduced the idea that thought should be characterised as a series of attitudes (such as *belief* or *desire*) towards language-like propositions (such as "it is raining" or "the sky is green"). Early work on propositional attitude reports, including that of Frege (1892/1980) and Russell (1910), focused primarily on elucidating a satisfactory semantics of propositional attitudes, but as we will see later on (in section 1.1.4), this characterisation eventually developed into a full-fledged, semi-empirical theory of cognition that to this day continues to influence the philosophical understanding of mind and cognition. In contrast to this characterisation I will argue that propositional attitude reports in fact constitute only a fraction of folk psychological discourse, and that equating folk psychology with propositional attitude reports gives a very narrow perspective on both cognition itself and the folk understanding of behaviour. Despite recent attempts to challenge this orthodoxy, it remains the dominant position in both philosophy and cognitive science, resulting in the oppositional nature of the debates over eliminative materialism (discussed in section 1.1.3), where it is assumed that folk psychology must either accurately describe the structure of cognition, or else be eliminated entirely.

1.1.2 – The Myth of Jones (1949-1953)

A crucial development in the formation of our contemporary understanding of folk psychology was Gilbert Ryle's influential attempt to analyse the use of mental state terms in natural language (see his 1949). Ryle advocated a form of philosophical behaviourism that maintained a principled scepticism towards the existence of unobservable mental states, which led him to characterise common-sense psychological states such as belief and desire as dispositions towards observable action rather than unobservable (and discrete) states of the human cognitive system. The philosophical behaviourist research program quickly ran into trouble, most notably as a result of high profile critiques by Chisholm (1957) and Geach (1957) on the philosophical side, and Chomsky (1959) on the side of cognitive science.

Nonetheless it presents an interesting alternative to the characterisation of folk psychology as attributing internal mental states, an alternative that Dennett (1987) arguably advocates a version of, and that I will return to in section 1.3.

Stich & Ravenscroft (1994: 450-3) trace the origins of folk psychology in the modern sense to Sellars' "myth of Jones" (Sellars 1956: 90-107), which describes how our fictional "Rylean" ancestors might have begun to ascribe propositional attitudes to one another, as well as to themselves. These ancestors begin with a language that refers only to external (i.e. non-mental) events and behaviours, before the eponymous genius, "Jones", develops a theory that relates these behaviours to postulated internal events (i.e. mental states). Sellars characterises these postulated internal events as propositional attitudes (*ibid:* 106), and quite aside from any further impact his argument might have had, this proto-functionalist analysis of mental states as propositional attitudes entered the philosophical mainstream.

DeVries (2006) explores how Sellars' myth has further influenced two modern philosophical projects, eliminativism and the theory of mind debate. I return to both below, in 1.1.3 and 1.1.5 respectively, but it is worth mentioning them here by way of illustrating the huge impact that Sellars has had on contemporary philosophy of mind and cognitive science. For eliminativism, Sellars provides the basis for the idea that folk psychology is a theory, and thus open to being judged according to the same standards as any other theory (*ibid*: 58). For the theory of mind debate, this idea is taken up by the theory-theorists, who argue that our understanding of other minds is theoretical in nature (*ibid*: 68). In both cases DeVries is sceptical as to the validity of interpreting Sellars as arguing that propositional attitudes are (simply) theoretical entities. Nonetheless, Sellars *has* been interpreted this way, and this interpretation has had a strong influence on how we now think about folk psychology.

1.1.3 – Folk Theory and Eliminativism (1966-1981)

Whilst Sellars might have been responsible for introducing the idea of propositional attitudes being theoretical entities, it was three papers by Lewis (1966, 1970, 1972)

that first explicitly formalised the idea of a folk psychological theory consisting of commonly accepted platitudes about mental states (see Ravenscroft 2010, sec. 3).

In the first paper (1966) Lewis describes the idea that mental states[2] should be characterised in terms of their causes and effects, and then identified with physical states that share these causes and effects. Furthermore, he allows that the causes and effects of mental states might be other mental states (*ibid* 21), thus forming a causal network of mental states. Lewis was not the first or only theorist to come up the general idea of treating mental states as causal networks, which appears in various forms of functionalism around the same time, including Smart's analytical functionalism (1959), Putnam's machine functionalism (1960, 1967), and Fodor's psycho-functionalism (1968). I will return to the development of this now classical approach to cognitive science in section 1.1.4.

In the second paper (1970) Lewis lays out a general structure for defining and interpreting theoretical terms, following earlier proposals by Ramsey (1931) and Carnap (1959, 1961, 1963, 1966). Ignoring some of the technicalities, he argues that theoretical terms (T-terms) can be defined in relation to observation terms (O-terms) that have "conventionally established standard interpretations" (Lewis 1970: 429). Thus T-terms are essentially shorthand for long strings of O-terms that uniquely pick out some feature of the world. Replacing any occurrence with an expanded O-term definition can eliminate T-terms, but there is no particular reason (short of outright reductionism) why we should want to do this, as it is simply more convenient to use the T-terms.

In the third paper (1972) he applies this account of theoretical terms to his earlier analysis of mental state terms in order to elucidate a common-sense or folk theory of psychology. The mental state terms posited by the folk are derived by treating our everyday talk about the mind "as a term-introducing scientific theory", and collecting/systematizing all (or most) of the platitudes contained in this theory (*ibid*: 256). Thus a mental state term is a kind of theoretical term, and we can treat

[2] In the paper he is actually talking specifically about "experiences", but the arguments apply to mental states more generally.

folk psychology as a whole as a primitive or naïve theory. Lewis explicitly relates this argument to Sellars' myth of Jones, and suggests that it provides a potential means by which to test the plausibility of that myth (*ibid*: 257). He also notes that according to his account, "the mental terms stand or fall together" (*ibid*: 258) – if they are part of an integrated and holistic theory, then the failure of any individual term to refer will prove fatal for the theory as a whole.

Churchland (1979, 1981) followed Lewis' lead in taking folk psychology to constitute a primitive theory, but pressed this final point by arguing that the theory was very likely to be false, and should therefore be replaced by a new theory drawn from "the conceptual framework of a completed neuroscience" (1981: 67). He called this position *eliminative materialism*, or *eliminativism* for short. His argument is that if folk psychology is genuinely a (proto-)scientific theory, then it ought to be held up to the same standards as any other scientific theory, which includes the potential for its eventual elimination or refinement. Churchland's eliminativism followed earlier projects by Feyerabend (1963) and Rorty (1965, see Leach & Tartaglia 2014 for an overview), but Churchland's argument is more explicit in targeting folk psychology as a propositional attitude invoking theory. Stich (1983) presents a slightly different form of eliminativism, arguing that *future* cognitive science is likely to contradict the theory embodied by folk psychology. In any case it is Churchland (and to a lesser extent Stich), rather than Feyerabend or Rorty, who has entered the popular philosophical imagination as the eliminativist *par excellence*.

Churchland's 1979 book and 1981 paper are also notable for another reason. It is here that the characterisation of folk psychology as propositional attitude psychology finds its clearest expression, along with a clear statement of the reasons for thinking that it constitutes a theory. Given that Lewis' set of papers were fairly technical, it is plausible to think that Churchland might have (somewhat ironically) ended up popularising the notion of a folk psychological theory, especially one consisting of propositional attitudes.

1.1.4 – Classical Cognitive Science (1968-1987)

Whilst folk psychology is rarely mentioned explicitly in cognitive science (outside of social cognition), it has played a role in formulating the conceptual scheme within which the scientific study of cognition is conducted. This is most apparent in what is sometimes called 'classical cognitive science', a research tradition that developed during the latter half of the 20th century and remains heavily influential today. The foundational assumption of this tradition is that cognition is basically a matter of computation, and that cognitive computation can be characterised as operations performed over what are essentially folk psychological states such as belief and desire.

Classical cognitive science dates back to the 1950s, when it emerged in opposition to the then-dominant behaviourist paradigm in psychology (see Thagard 2012: sec. 1). Chomsky's extremely critical review of Skinner's *Verbal Behaviour* was a crucial turning point in this respect. Right from the beginning this research program was driven by the search for what Putnam (1967) characterised as functional states. These were often assumed to be isomorphic with the apparent posits of folk psychology, as made most explicit in Fodor's language of thought hypothesis (see his 1975).[3] Here Fodor argued for the logical necessity of an internally represented language, *mentalese*, which would encode the propositional attitudes and allow for their systematic manipulation in cognitive processing. Thus Fodor combined Lewis' formalisation of folk psychological structure with Putnam's functionalism to explicate the newly emerging paradigm that would implicitly guide cognitive science for at least the next decade (cf. Pickering & Chater 1995: 313; Piccinini 2004a, 2004b). Fodor also makes an explicit claim about the universality of propositional attitude psychology (or at least beliefs and desires), even going so far as to deny that there are any conceivable alternatives (1987: 132). This move by Fodor enshrined the conflation of folk psychology with propositional attitude

[3] Strictly speaking there is something of an anachronism here, as folk psychology only explicitly entered the philosophical discourse *after* the emergence of classical cognitive science. However, prior to this something like propositional attitudes were already being invoked with the idea of discrete internal states that might guide behaviour.

psychology, and committed classical cognitive science to one very particular way of understanding cognition.

1.1.5 – Social Cognition (1978-present day)

Social cognition is the specific study of social interaction from the perspective of psychology and cognitive science. Premack & Woodruff (1987) established the problem of other minds as a field of legitimate psychological interest, rather than as a question of merely philosophical interest. The primary distinction here is between asking how we can know whether or not other people have minds at all (a primarily philosophical question), and asking the more specific question of how it is that we can understand their mental lives, assuming as a starting point that they do in fact have one (a question that both philosophy and cognitive science can contribute to). Since the publication of this paper there have been a number of interesting developments, both theoretical and experimental. I will summarise the historically relevant aspects below, before discussing social cognition in more detail later in this chapter.

Initially the dominant theory in social cognition was theory-theory (cf. Morton 1980), which postulates the tacit use of a literal theory of how minds work as the primary cognitive mechanism for understanding other minds. This theory is normally fleshed out in terms of propositional attitudes, and thus embodies the identification of folk psychology with propositional attitude psychology that I introduced in the previous section. As Baker (1999a) notes, this has led to some confusion in the philosophical literature. It is often unclear what kind of a theory mindreading requires (cf. Botterill 1996, Lavelle 2012) or even whether it makes sense to characterise common sense psychology as a theory at all (Baker 1999b). Perhaps as a result of this, there have been several attempts to account for our understanding of other minds in non-theoretical terms, most notably including versions of simulation theory (e.g. Gordon 1986; Heal 1986; Goldman 1989, 2006), and interaction theory (e.g. De Jaegher & Di Paolo 2007; Gallagher 2008a, 2012). Each of these accounts distinguishes the mechanisms by which we understand other

minds from our explicit verbal reports about mental states and behaviour, thereby detaching, at least partially, social cognition from folk psychology. More recently it has been suggested that explicit folk psychological discourse might even have a very different purpose to the mechanisms that we use to understand other minds. A number of writers (e.g. McGeer 2007; Hutto 2008; Zawidzki 2008, 2013; Andrews 2008, 2015) have argued that folk psychology has a social or normative role, quite distinct from the epistemological role played by social cognition. I will return to the broader implications of these arguments later in this chapter, and again in chapter 3.

1.1.6 – Connectionism and Embodiment (1986-present day)

Since the late 1980s there have been a number of trends in cognitive science that have called into question various aspects of the folk psychological framework. The first of these was the move towards connectionist models of neural processing (usually dated to Rumelhart & McClelland 1986), which envisioned cognition as a process that takes place across wide and parallel neuronal structures, rather than via serial computations at a higher level of abstraction. This brought with it the possible elimination of the symbolic level of processing, or even of folk psychology entirely (see Ramsey, Stich, & Garon 1991; see Fodor & Pylyshyn 1988 for a defence of the classical account). Connectionism received an early endorsement by Churchland, and also became associated with other eliminative materialists such as Stich. The debate over the implications of connectionism remains controversial, but it undoubtedly changed the way that we think about the interaction between different levels of cognitive processing, and what this means for folk psychology (see e.g. Clark 1990, O'Brien 1991, Botterill 1994; see Clark & Millican 1999 for an overview of key issues). One important outcome was the idea, now broadly accepted, that folk psychological or otherwise 'high level' mental states could be implemented with 'lower level', sub-symbolic processing, such as that proposed by connectionism.

Following connectionism, the early 1990s to the early 2000s saw the emergence of a broad paradigm known as embodied cognition (see Shapiro 2010 for an overview). This consists of a number of distinct research programmes, but

essentially emphasises the role of the body and/or the environment in cognitive processing. These programmes are sometimes associated with a non-classical interpretation of folk psychology, in that they often downplay the role of abstract, conceptual, or serial reasoning (see e.g. Brooks 1991; Beer 1995; Van Gelder 1995; Garzon 2008). More modestly, it seems likely that embodied approaches to the study of cognition would at least challenge the disembodied interpretation of folk psychological explanation that is associated with the classical approach. This is the assumption that folk psychological discourse deals mostly in the attribution of mental states alone, rather than with the prediction of behaviour alongside more complex capacities such as narrative competency (I discuss all of this in more detail later in the chapter).

Sometimes associated with embodied cognition, but in many ways importantly distinct, is the hypothesis of *extended* cognition associated primarily with the work of Andy Clark (Clark & Chalmers 1998; Clark 2008; Menary 2010). This proposes that cognitive processes might literally extend into the environment, which initially seems in radical opposition to our folk conceptions of the mind. However, as I will explore in chapter 2, it is possible that the brain-bound conception of the mind is partly an artefact of the specific European-American scientific culture from which cognitive science emerged. If this is the case then extended cognition might in fact be less radical than it first appears.

In his more recent work Clark has drawn attention to a growing cluster of theories that model the brain as a Bayesian prediction machine (Hohwy 2013; Clark 2013, 2016). Clark has written that predictive processing "may one day deliver a better understanding even of our own agent-level experience than that afforded by the basic framework of 'folk psychology'" (2013: 17, repeated in Clark 2016: 82). I will explore the relationship between predictive processing and folk psychology in chapter 4.

1.2 – Folk Psychology as Social Cognition

At the beginning of this chapter I distinguished between two primary ways that the term 'folk psychology' can be used, referring either to folk psychological discourse or to the cognitive mechanisms that enable social cognition. In the rest of this chapter I will describe in more detail what each consists of, and discuss several on-going debates about the nature of each kind of folk psychology. This first section will focus on folk psychology understood as mechanisms that enable social cognition. It is important to distinguish between the implicit understanding of the mind imparted by these mechanisms, and the explicit understanding contained within folk psychological discourse (I discuss this distinction in more detail in 2.1.1). There is likely to be some overlap between folk psychological discourse and social cognition, especially when it comes to the content of attributed mental states and the regulative function that I am calling mindshaping (see 1.3.5, 1.4, and chapter 3 for further discussion).

Social cognition, in the broadest possible sense, refers to the cognitive mechanisms and processes that enable interpersonal understanding. This includes language processing, joint attention, and many other capacities, but the study of social cognition has typically focused on one core aspect: the attribution of mental states to others, sometimes known as mindreading or theory of other minds (although this latter term begs the question somewhat against accounts that do not posit a theory). The exclusive focus on mindreading in social cognition has recently been criticised (see e.g. Hutto, Southgate, & Schwenkler 2011; Zawidzki 2013), and I will return to these criticisms towards the end of the section, but it remains the core area of study to this day.

1.2.1 – Theory-theory and the false belief task

As I mentioned in the previous section, the formal study of social cognition began in the 1970s, after Premack & Woodruff (1978) published a *Behavioral and Brain Sciences* target article that argued that chimpanzees might have a "theory of mind" nearly as advanced as our own. This target article not only introduced the term

'theory of mind' as a characterisation of interpersonal understanding, it also described an experimental framework for testing whether or not chimpanzees do in fact possess such a theoretical understanding of other minds. The debate about whether or not chimpanzees have a theory of mind continues to this day, but Premack & Woodruff's paper was also instrumental in kick-starting the study of *human* social cognition, as researchers began thinking about how to implement similar experiments with human subjects. These developments were partially prompted by philosophical commentaries written by Dennett (1978), Bennett (1978), and Harman (1978), each of which "independently suggested that a proper test of a creature's possession of the belief concept would involve the determination of its ability to impute false belief" (Goldman 2006: 11).

By far and away the most successful and influential of the experimental paradigms inspired by Pemack & Woodruff was the false-belief task, first conducted by Wimmer & Perner (1983), and subsequently developed by Baron-Cohen, Leslie, & Frith (1985), who coined the name "Sally-Anne task" by which it is commonly known (in reference to the dolls used in their version of the task). The basic structure of the task is that the participants (typically children) are presented with a scenario in which two actors (either dolls or humans) interact. One actor hides an object and then leaves the scene. The second actor then enters, hides the object somewhere else, and leaves. Finally the first actor returns, and the participant is asked where they think the actor will look for it. If they correctly identify that the actor will look where they originally hid the object, rather than where it actually is, then they are said to have an understanding of false belief, and by extension a fully developed theory of other minds. There are many variations on this task, the most well known of which involves tracking looking time (as a proxy for expectation) rather than (or as well as) verbal responses (e.g. Onishi & Baillargeon 2005). The development and evolution of this task has been extremely important to the study of social cognition, as it established a broadly accepted measure of knowledge of other minds, and allowed for subtle experimental manipulation in a way that previous studies of social interaction had not. For instance, by varying the task structure to involve food rather

than toys, and conspecifics or experimenters rather than dolls, it is possible to probe the social cognitive competency of non-human animals such as chimpanzees and orangutans (see Heyes 1998 for an overview).

The aim of Baron-Cohen, Leslie, & Frith's false-belief task was to determine the extent to which development of a theory of other minds was impaired in children with autism. Subsequent applications of the experimental design were used to put together a more general understanding of the developmental trajectory of social cognition in human children. One early finding was that children below the age of around 4 typically failed to pass the test, and so it was argued that a theory of other minds did not develop until the age of 4 or 5. However more recent versions of the task (see e.g. Onishi & Baillargeon 2005), based on using looking time as a proxy for expectation, have demonstrated that children as young as approximately 15 months seem to expect the actor to look in the wrong location. The debate about how to interpret these results in still on-going, and I will return to it in chapter 4, as it serves as an interesting illustration of how folk intuitions can potentially disrupt psychological experimentation.

The main theoretical outcome of the false belief task paradigm was the emergence of what has become the establishment position in social cognition: the idea that interpersonal understanding is based on an implicit theory of other minds, sometimes known as the 'theory-theory'. The theory-theory also draws on the philosophical tradition mentioned in the previous section, taking the idea, introduced by Sellars and developed by Lewis and Churchland, that we can treat the common sense understanding of how other people behave as an either implicit or explicit theory of how their minds work. There are two main versions of the theory-theory, one that claims we have an innate theory of mind module (Leslie 1994, 2000) and one that claims that children develop a theory of other minds in a proto-scientific manner (Gopnik & Wellman 1992).

1.2.2 – Simulation theory and mirror neurons

The earliest proposed alternative to the theory-theory was the idea that rather than theorising about other people's mental lives, we could just put ourselves into their shoes, so to speak, by using our own cognitive system to simulate their mental processes based on the situation that we see them in. This alternative proposal is known as simulation theory. It was first formulated by Gordon (1986), who proposed that we predict other people's behaviour in the same way that we predict our own. This can of course be interpreted in different ways (Sellars, for instance, suggested that introspection might consist of self-directed folk psychological theorising), but Gordon took it to mean that we might interpret the behaviour of others by pretending that we were in the same situation as them, and then simply "speak our mind" about what we felt they would do (*ibid*: 160). Heal (1986) made a simultaneous argument for a similar approach based on "replication" rather than "simulation".

These proposals were then developed by empirical research carried out throughout the 1990s, culminating in the discovery by Vittorio Gallese, and others in working in his lab, of so-called "mirror neurons" in rhesus macaques (see Di Pellegrino *et al* 1992, and Rizzolatti *et al* 1996). This cluster of neurons in the inferior frontal gyrus and inferior parietal lobe fire both when an activity is carried out *and* when the macaque observes someone else carrying out the same activity, leading to the proposal that the mirror neuron system might be essential to social cognitive simulation. Mirror neurons have only been inferred to exist in humans, as no direct observation via single cell recording has been carried out in human subjects, but Gallese and others have argued at great length for the fundamental role played by the mirror neuron system in social cognition. Whilst the simulation theory is not necessarily committed to the existence of a human mirror neuron system, in practice many proponents of it have bought in to Gallese's approach (see especially Goldman 2006).

There have been numerous criticisms of the simulation theory, including early defences of theory-theory by Stich & Nichols (1992), Gopnik & Wellman

(1992), and Perner & Howes (1992), although Perner has since come to defend a hybrid approach (which I discuss in the next section). One common criticism is that whilst simulation might form a part of how we understand other minds, it cannot do all the work by itself, as it needs to be embedded in a theory that tells us how and when to simulate the mind of another, and also tells us which aspects of theory in question are relevant to the simulation. The mirror neuron theory in particular has also come under attack for lacking empirical support in humans (see e.g. Hickok 2009).

1.2.3 – Alternative theories

In recent years the debate between simulation theory and theory-theory has reached something of an impasse, leading to the development of a range of alternative theories. One type of alternative theory, known as hybrid theories, attempt to agree on a middle ground between the two theories. What most of these hybrid accounts have in common is the idea that whilst simulation may play an important role in mindreading, it requires the support of a theory in order to be correctly applied (see e.g. Botterill 1996; Goldman 2006; Carruthers 2011; Stich and Nichols 2003). The theoretical component of the hybrid accounts is also used to account for systematic errors and to explain how we can successfully predict the behaviour of people in situations that we have never been in before. Insofar as the debate between theory-theory and simulation theory has reached an impasse, hybrid accounts seem like a sensible way forward, although later in this section I will suggest that they still miss some crucial aspects of social cognition.

Another type of alternative theory are those that argue for the existence of two distinct systems for social cognition, one fast and operating in response to observed behaviour, the other slow and theoretically mediated (see e.g. Apperly & Butterfill 2009). These accounts were originally inspired by the so-called "developmental paradox of false belief understanding" (De Bruin & Newen 2014), which is the strange result that infants are able to pass a non-verbal version of the

false belief task well before they can pass the standard, elicited response version.[4] According to the two systems account, the non-verbal task is processed by system one (fast), which develops earlier in life than system two (slow), but does not allow for explicit verbal reports. Hybrid theories of this type are not exactly the same as a combination of simulation theory and theory-theory, but they share some important properties, such as positing the existence of two distinct mechanisms for social cognition. Two systems proposals are also somewhat orthogonal to the theory-theory/simulation theory debate, as one could combine either approach (or a hybrid approach) with a two systems architecture. I will discuss these theories in more detail in section 2.4.2.

1.2.4 – Interaction theory and direct perception

A relative latecomer to the social cognition debates, although pre-empting Apperly & Butterfill, are the various interaction theories proposed by the likes of Shaun Gallagher (2008a, 2012), De Jaegher & Di Paolo (2007), and Dan Hutto (2008, 2009). Each of these theories differs somewhat in emphasis, but what they all have in common is a commitment to the idea that social cognition, as it has traditionally been studied, has focused too much attention on internal processing (either via theory or simulation) at the expense of external interaction with other social agents. Thus Gallagher argues that social cognitive development consists of three distinct stages: primary intersubjectivity, secondary intersubjectivity,[5] and narrative competency (Gallagher 2008a), which are defined as follows:

1. *Primary Intersubjectivity:* Newborn infants respond to voices, faces, and movement, and are soon able to engage in interactive imitation and response.

[4] In a typical non-verbal false belief task the subject will be placed in front of an eye-tracker, which will keep track of how long they fixate on any given scene, with length of fixation being used as a proxy for how unexpected that scene was. So if they look longer during the scene where Sally looks in the right place despite having a false belief, then it is inferred that the subject does have the capacity to attribute false beliefs, explaining why they were surprised.

[5] The terms primary and secondary intersubjectivity, and the associated account of social cognitive development, originates in the work of Colwyn Trevarthen (see especially Trevarthen 1979; Trevarthen & Hubley 1978).

By 12 months they are capable of what Gallagher calls "non-mentalistic, perceptually-based embodied understanding of the intentions and dispositions of other persons" (2008a: 166).

2. *Secondary Intersubjectivity*: After 12 months infants begin to develop an awareness of pragmatic context, along with a capacity for shared/joint attention. In some direct sense they "are able to see bodily movements as expressive of emotion, and as goal directed intentional action" (*ibid*).

3. *Narrative Competency:* By the age of 4 (but from as early as 2) children are able to situate their interaction with others within a detailed narrative framework that provides a capacity for empathic and novel understanding (Gallagher 2012: 16-8). This framework is distinct from a theory in the sense that it is not reliant on folk psychological or propositional mentalising (*ibid*: 18-9).

Adult social cognition, according to Gallagher, involves capacities drawn from each developmental stage, but only rarely consists of explicit theorising or simulation. Similarly, De Jaegher & Di Paolo (2007) argue that social cognition primarily consists of "participatory sense-making", involving both explicit narratives and embodied intersubjectivity. I will return to these suggestions later in this chapter, and propose that regardless of their status as basic social cognitive mechanisms, these explicit forms of intersubjective understanding will inevitably form an important part of folk psychological discourse considered more broadly.

A distinct challenge emerging out of the interaction theory camp is to argue that both theory-theory and simulation theory incorrectly characterise social cognition as an indirect process, mediated by either a theory or a simulation. Gallagher (2008b) argues that for both phenomenological and theoretical reasons this cannot be possible, and that social cognitive perception is in fact direct and immediate. Gallagher draws on Gibson's ecological theory of perception (see e.g.

Gibson 1966, 1979), whilst theory-theory is firmly grounded in more mainstream inferential theories of perception (see e.g. Fodor & Pylyshyn 1981). As such, the debate over direct perception may just reflect a more general theoretical disagreement, rather than being a dispute about social cognitive perception in particular (cf. Dewhurst, ms). In any case, Lavelle (2012) has responded to Gallagher's criticism by proposing a way in which theory-theory can be made compatible with direct perception. This suggests a general trend towards a growing consensus approach, which I discuss in more detail at the end of this section.

1.2.5 – Mindshaping and social regulation

There is one final alternative to the classical debate between theory-theory and simulation theory that I wish to mention, especially as it is one that I think plays an important role in developing an alternative, non-epistemic role for folk psychological discourse. This is the idea that social cognition might involve more than prediction and explanation, that it might also play a normative role in constraining our behaviour and cognition. A version of this idea was first proposed in the modern context by McGeer (2007), although it has precursors in Dennett (1989) and Morton (1980). It was then developed independently by Zawidzki (2013), who builds on the idea of "mindshaping" introduced by Mameli (2001). I discuss both proposals in more detail in section 1.3, and again in chapter 3, which is dedicated to the topic of mindshaping and social regulation.

McGeer (2007) claims that folk psychology, even under more traditional accounts, contains what she calls a "normative core", i.e. the assumption that folk psychological attributions can somehow 'make sense' of the behaviour that they are attributed to. In order to do this, folk psychology must appeal to some minimal norms of rationality (*ibid*: 140-5). From here she argues that it is no great leap to see folk psychological attributions as performing a further regulative role, where our attributions also carry social expectations that those to whom they are attributed typically end up trying to fulfil.

The normative role that McGeer describes for folk psychological attributions is just one component of the broader framework articulated by Zawidzki (2013). He describes social cognition as a complex relationship between mindreading, mindshaping, co-operation, and symbolic communication. According to his account each of these capacities is crucial to what he calls the 'human sociocognitive syndrome', and yet only the first has really received much critical attention from within philosophy. He focuses on explicating a structured framework for studying the second, and advocates a shift from the "mindreading-as-linchpin" hypotheses, which emphasises mindreading both onto- and phylogenetically, to the "mindshaping-as-linchpin" hypothesis. I will discuss Zawidzki's proposals in more detail in chapter 3, but introduce them here to give a full picture of the current state of the art in social cognition.

1.2.6 – Towards a unified account of social cognition

Although it is not the primary aim of this thesis, I would like to suggest in passing that there seems to be a general move towards a unified account in social cognition, based on a growing consensus that purely theoretical disputes should be replaced with empirically verifiable hybrid approaches, or left by the wayside if no empirical resolution is forthcoming. Something like this has already occurred in the case of theory-theory/simulation theory hybrids, and also in the disputes of behaviour reading vs. mindreading in primate social cognition. Lavelle's (2012) proposal for reconciling direct perception with theoretical inference marks another such attempted reconciliation of a classic debate. Zawidzki's novel proposal that social cognition should emphasise mindshaping over mindreading leaves room for more traditional work that focuses exclusively on mindreading, and there is no in-principle reason why a rich account of mindshaping should not be compatible with a hybrid approach to mindreading that allows for direct perception. We seem to be entering an era where social cognition is beginning to develop into a mature scientific discipline, with a relatively stable core theory that will allow for more interesting and exploratory research into specific issues such as joint attention, the relationship

between mindshaping and mindreading, and the recursive influence of language and culture on social interaction. With that in mind I will move on from social cognition as such, and look in more detail at the folk psychological discourse that it produces (and is perhaps part-constituted by).

1.3 – Folk Psychology as Folk Discourse

In the first two sections of this chapter I looked at the historical usage of the term folk psychology, and at the historical and contemporary study of social cognition, which is sometimes (potentially misleadingly) described as the study of folk psychology. In this penultimate section I want to argue for a novel characterisation of folk psychology as an explicit folk discourse or practice, one that encompasses multiple different components including traditional propositional attitude psychology, mental state attribution more generally, behaviour reading, social regulation, and narrative competency. This explicit discourse is importantly distinct from the implicit social cognitive mechanisms that I described in the previous section, although the latter play an important role in generating the former, and may in turn be constrained and influenced in interesting ways by the folk discourse. In this section I will describe what I mean by a folk psychological discourse, and discuss each of the elements that it is composed of. In the final section that follows I will say something about how I envisage the relationship between folk psychological discourse and social cognition.

1.3.1 – Folk psychological discourse

By describing folk psychology as a discourse, I hope to encourage a more open-minded and pluralistic approach, in which it becomes conceivable that our everyday descriptions of behaviour could consist of more than just theoretically motivated ascriptions of propositional attitudes. I also want to clearly distinguish folk psychology in this broader sense from the more traditional accounts that I have described in previous sections. I will go on to argue that we also describe and predict behaviour in non-mentalistic terms, situate our descriptions and predictions in an on-

going narrative structure, and make explicit normative judgements about how one *ought* to behave – where "ought" is understood as carrying both ethical and rational weight. I do not want to give any formal definition of what a discourse is, and I certainly do not have in mind the more critical sense of a socio-political discourse that comes out of the works of Foucault and other social theorists (although studying folk psychological discourse in this latter sense might well constitute an interesting project in its own right). The rest of this section will focus on discussing in more detail the various facets of folk psychological discourse, but first I will say a little about how I see them all fitting together.

In our everyday lives, we are often able to make predictions about how someone is likely to behave, both in the short and long term. We might make such predictions explicitly, and can typically verbalise them if asked (although we might just keep them to ourselves). Or we might do this implicitly, via some combination of the processes and mechanisms described in the previous section. For my purposes it doesn't matter: any basic behavioural prediction of this kind qualifies as what I am calling 'behaviour reading'. Sometimes we may also attribute mental states when predicting behaviour, although more typically such ascriptions are used in an attempt to explain or justify either past or future behaviour. Another type of explanation or justification is narrative competency, which is distinct from either behavioural or mentalistic language in that it embeds a person's actions in a wider narrative, and may not attribute any agency to their actions in particular. All three of the above kinds of folk psychological discourse (i.e. behaviour reading, mental state attribution, and narrative competency) can also be used to impose normative constraints on behaviour, i.e. by using them in an imperative rather than descriptive mode. I discuss each of these four categories of folk psychological discourse in more detail below.

1.3.2 – Behaviour reading

This is the most basic category of folk psychological discourse, and it overlaps somewhat with the implicit social cognitive mechanisms that I discussed in the last section. Regardless of how our behavioural predictions are generated, it is

undeniably true that people are in general relatively competent at predicting the future behaviour of their conspecifics – at least under normal circumstances. At a very basic level, we are able to avoid bumping into strangers on the street and can make use of physical cues to understand what someone is about to do, or what they expect us to do. Below I will consider a couple of more complex examples that demonstrate the kind of things that we can achieve using behavioural predictions alone.

When I see my colleague get up from her desk and head empty-handed towards the fridge that stands in the corner of our office I can safely predict that she will probably open it and take something out. Of course, predictions of this kind only go so far – without any additional information I probably couldn't tell what she was going to get out of the fridge, although once I knew what she had got out I could probably predict what she was going to do with it. The additional information required to predict what she might get out is precisely what is provided by the other components of folk psychological discourse. For example, if I had seen her put some chocolate in there earlier in the day, and if she had just told me that she fancied a snack, I might be able to successfully predict not only that she would open the door, but that she might take out the chocolate, break off a piece, and eat it. If I was further aware of her kindly nature, and that she knew I liked chocolate, I might predict that she would offer me some as well. So behavioural predictions that go beyond very simple and immediate circumstances seem to typically require further, non-behavioural information.

Having said that, there is quite a lot that we can achieve with behavioural predictions alone, provided that we are familiar with the situation and/or person in question. Consider another example. This time, I am observing a group of people who are engaged in a team sport. Based on my knowledge of the sport in question, I can predict what the players are likely to do, without needing to attribute any kind of hidden mental states to them. My predictions will undoubtedly not be perfect, but will certainly go beyond what I could achieve if I was not familiar with the sport. Add in a little extra information about each player, such as how they have tended to

play in the past, and my predictions will once again become more reliable. At this point it perhaps becomes an open question whether what I am doing should be described as behavioural prediction or explicit theorising, as my information about each player's previous performance could either be something I am explicitly aware of, or something I only know tacitly. The precise answer to this question does not matter for my purposes – it is sufficient that at least some of the time I might predict the behaviour of these sports players without attributing mental states to them.[6]

As the previous example demonstrated, we can improve on our basic capacity for behavioural prediction by engaging in explicit theorising. Rather than just predicting future behaviour on the basis of current behaviour, I can supplement my prediction with a model of the kind of situation that I am observing, and how it normally plays out. Note that this is distinct from the kind of implicit theory proposed by the theory-theory, which I do not have explicit access to, and which is based on a set of assumptions about how the mind works, rather than the situation which I am in.[7] Explicit theorising of the kind that I mention here can be purely behavioural, but it can also posit hidden mental states, or be expressed in terms of a narrative structure. I expand on each of these possibilities below.

1.3.3 – Mental state attribution

As we saw in the previous sub-section, there is a very thin line between exclusively behavioural predictions and more complex attributions of mental states as hidden causes of behaviour. The latter are what have typically been emphasised in previous philosophical discussions of folk psychology, normally under the more specific guise of propositional attitude attributions. I think it is important to distinguish between

---

[6] Botterill (1996: 107) notes that we could predict the outcome of a football game without having access to a 'theory of football', but I want to go one step further and say that we can sometimes predict the behaviour of a football *player* without having access to, or without having to use, a theory of *mind*.

[7] Although a theory theorist could argue that we do in fact have access to the content of our theory of mind, and that mental state attribution involves explicit theorizing, in which case my explicit theorizing about the behaviour of this sports team might be an example of theory-theory in action. This version of theory-theory is not very popular, however, and seems to suffer from a number of phenomenological and conceptual inconsistencies (see e.g. Gallagher 2008b).

mental state attributions in general and the particular case of propositional attitude ascriptions, if only to make room for the in-principle possibility that there could be non-propositional mental states attributed by the folk, such as character traits like kindness. With that in mind, let's consider what kind of mental states are in fact attributed in folk psychological discourse, and additionally what function such attributions might serve.

When giving behavioural descriptions and predictions of the kind I sketched out in the previous sub-section, it is extremely natural to supplement the description with additional mentalistic language, to the point where not doing so can in fact feel somewhat artificial. Consider again my prediction of what my colleague will do when she stands up from her desk and walks towards the fridge. Based only on behavioural assumptions, I can predict that she will open it and take something out, and perhaps even predict what she will take out if I saw her put something there earlier, or if she only ever uses the fridge to store one item, but my predictions immediately become much more powerful if I have access to mentalistic data. Now I can predict that she will take out some chocolate and eat, because I know that she is hungry, and that she believes there to be some chocolate in the fridge. I can also predict that she will offer me some, because I know that she is kind, and I know that she knows that I am hungry. This is only a very simple case, but immediately the complexity of both the predictions and explanations begins to increase, especially once we get into recursive attributions ("I know that she knows that…").

Note that we have at least two kinds of attributions in the above vignette, at least prior to further analysis. We have propositional attitude attributions: "she believes *x*" and "she knows *x*"[8]. We also have something like emotional or dispositional attributions: "she is hungry" and "she is kind". Whilst these can easily be reinterpreted as propositional attitude attributions – "she desires food" and "she wants to please others" – I think that doing so mischaracterises the nature of the folk

[8] The predicate "knows" must here be treated as a naïve folk psychological attribution – nothing of any epistemological significance is intended, and "knows" could perhaps be reinterpreted as a strong version of "believes" in almost every case.

discourse, as we typically interpret dispositional attributions as having a wider remit than propositional attitude attributions. "She is hungry" implies not only that she desires food, but also that she might be somewhat irritable, and that she might use food-related examples when making philosophical arguments, for instance. In a sense these kinds of attributions constitute a basic narrative, which I will discuss in more detail in the next subsection.

Note also that I slipped into folk psychological discourse in order to justify *my own* predictions, such as when I stated that I could predict her behaviour "because **I know** that she is hungry". This points towards not only the prevalence of folk psychological discourse, but also a justificatory function that it performs in addition to prediction and explanation. One is implicitly required to provide folk psychological justifications for one's behaviour in all kinds of situations, and we are typically able to give such justifications even if they are not normally elicited. Later in this section I will argue that even if these justifications are typically *ad hoc* and unreliable, they provide constraints on our future behaviour that can end up turning them into self-fulfilling prophecies.

So, we have at least two kinds of mental state attributions, propositional attitude and dispositional, both of which appear to greatly expand the predictive success and explanatory power of folk psychological discourse. Whilst this explains some of the historical focus on propositional attitude attributions, in the rest of this section I will argue that there are other, equally powerful components of folk psychological discourse.

1.3.4 – Narrative competency

As we have seen in the previous two sub-sections, there is a natural progression from behaviour prediction, to mental state attribution, culminating in full-blown folk psychological narratives. Whilst I can predict my colleague's behaviour by engaging in crude behaviourism or by attributing mental states, it is often easier (and perhaps even more natural) to simply situate her actions in an on-going narrative structure, one that I have built up over the weeks, months, and years that I have known her.

This narrative allows for predictions, as if I am familiar with the narrative then I know what comes next, but it also provides a contextual justification for her behaviour (the importance of which I will turn to in the next subsection). Bruner (1990) first introduced the notion of a folk psychological narrative, but it has since been developed extensively by Hutto (2008).

We do not only acquire folk psychological narratives via direct observation of conspecifics – if this were the case then this capacity would arguably be no more than a complex version of behaviourism. As part of our social cognitive and folk psychological development, we learn all kinds of narratives, either by explicitly being taught them in the form of fictional stories, or through implicit observation of generalizable everyday situations.

It is important to recognise that narrative competency cannot be all there is to folk psychological discourse. Hutto notes that narratives seem to be inherently reliant on "having a grasp of the core propositional attitudes" (2008: 129). Similarly, Gallagher's alternative account of social cognition relies on the acquisition of what he calls primary and secondary intersubjectivity prior to narrative competency (2008a). Whilst according to my taxonomy these are more properly understood as social cognitive rather than folk psychological mechanisms, because they take place below the level of explicit awareness, they do both include elements of behaviour reading, and it seems right to say that without some kind of basic understanding of people as predictable agents, it would be hard to make sense of narratives that included them as characters. Narrative competency is best thought of as a continuation and complexification of behaviour reading and mental state ascription, buttressed with socio-cultural schemas, rather than a totally different way of understanding behaviour.

1.3.5 – Normative constraints

The final component of folk psychological discourse that I will consider in this section is the normative or regulative pressure that this discourse can exert on us. Morton (1980) first articulated the suggestion that folk psychology might be partially

normative, and Mameli (2001), McGeer (2007), Zawidzki (2013), and Andrews (2015) have all explored it more recently. I will discuss this idea and its implications in more detail in chapter 3, but for the time being I focus on just giving a rough impression of how it fits into the overall picture of folk psychological discourse.

Folk psychological discourse can be considered to be playing a regulative role whenever it causes us to adjust our behaviour in some way. Zawidzki lists several different forms that this can take, "including imitation, pedagogy, norm cognition and enforcement, and language based regulative frameworks, like self- and group-constituting narratives" (2013: 29). Note that mindshaping, as Zawidzki calls it, spans the whole range of folk psychological discourse, from "self- and group-constituting narratives" right down to the basic, perhaps pre-folk psychological, imitation of the behaviour of conspecifics. One particularly interesting case that Zawidzki explores in some detail is the way in which our explicit attributions of mental states to ourselves and to others ends up becoming a self-fulfilling prophecy, as we consequently feel social pressure to conform to those attributions and thus maintain at least a kind of surface level consistency. For example, if someone attributes to me the belief that it is raining, they are not only making an epistemic claim, but also exerting social pressure on me to perform certain actions (putting on a raincoat, carrying an umbrella, etc.) at risk of otherwise looking either irrational or contrary, or else making *them* look foolish, which comes with its own social costs.

Another kind of normative constraint is imposed when we try to justify our behaviour or the behaviour of others. Whilst such justifications can often be *ad hoc,* and may not closely match our actual reasons for doing something (assuming such a reason even existed in the first place), they do help us make sense of our behaviour as situated in a narrative, and they also exert an influence on our future behaviour as we try and keep this narrative consistent. Hutto suggests that folk psychological narratives are in fact primarily invoked "to make sense of […] seemingly aberrant actions" (2008: 37; also McGeer 2007). He goes on:

> Bruner is right that rather than merely providing a framework for disinterested prediction and behaviour, narratives – and especially folk psychological narratives – work to regulate our actions; as such they are "instruments of culture": they summarize "not simply how things are but (often implicitly) how they should be." (Hutto 2008: 37, quoting Bruner [1990: 40])

Andrews (2015) gives the helpful example of being late for a meeting, and explaining that you were late because of bad traffic. This story justifies your late arrival, but at the same time also exerts a normative pressure on you to consider alternative forms of transport in the future – there's only so many times that the excuse will work before your colleagues grow tired of it. I explore these mechanisms in much more detail in chapter 3.

1.3.6 – Summary of folk psychological discourse

I have described folk psychological discourse as being a complex phenomenon composed of (at least) four somewhat distinct capacities. At perhaps the most basic end, we are able to make successful short-term behaviour predictions, such as when we side step someone to avoid bumping into them on the street. More complex capacities include the explicit attribution of mental states, the invocation of narratives to explain behaviour, and the establishment of normative constraints that help structure our social environments.

These capacities are obviously closely related to one another, and distinguishing between them in this way is inevitably going to feel somewhat artificial. Mental state attributions can be used in order to help predict behaviour, and narratives are an important component of the normative constraints that we impose through folk psychological discourse. Nonetheless they are distinct enough that we can make sense of talking about them separately, even if this is often a simplification of the actual state of affairs.

1.4 – Folk Psychological Discourse and Social Cognitive Mechanisms

So far in this chapter I have presented a history of the term folk psychology, and explored two distinct ways in which it can be used: to refer to social cognitive mechanisms and to refer to folk psychological discourse, the latter being a personal level description of how we actually engage with one another in a social context, whilst the former describes the sub-personal mechanisms that make social interaction possible. The previous sections have surveyed the current state of play with respect to both these uses of 'folk psychology', and provided some brief suggestions about the direction in which I think debates in each area should be headed. In this final section I will consider how these two phenomena interact with one another.

As I have already noted, the distinction between social cognitive mechanisms and folk psychological discourse is to some extent an artificial one. Depending on which account of social cognition you think is correct, you might think of the folk discourse as simply being an articulation of the implicit knowledge embodied in our theory of other minds. In this case the two would be more or less identical, although perhaps some components of the folk discourse, such as narrative competency, would be seen as a culturally mediated addition to the core theoretical structure. Alternatively one could see the entirety of the folk discourse as an additional layer built on top of our basic capacity for social interaction – this is something like the position expressed by Gallagher (2008a), although one could remain more neutral with regard to what form the social cognitive mechanisms take.

There are further important questions to be asked about the interaction between social cognitive mechanisms, which I take to be relatively implicit and at least to some extent universal, and folk psychological discourse, which is highly culturally mediated and, as we will see in chapter 2, nowhere near as universal as traditionally assumed. To what extent is the folk discourse constrained by our actual capacity for social cognition? To what extent are social cognitive mechanisms influenced by the vagaries of folk psychological discourse? Whilst answering these questions is not the main aim of this thesis, I will briefly say a few words about each.

Our capacity for successfully predicting behaviour is a product of social cognitive mechanisms, although as I suggested in the previous section, it is probably also facilitated by normative constraints imposed by folk psychological discourse. This makes behavioural predictions an interesting borderline case – whilst we sometimes make explicit reference to them in our folk discourse, they are more often at work 'behind the scenes', where we might even lack explicit awareness of them. It is typically only when someone's behaviour is odd or unexpected, or when someone else requests guidance, that we engage in explicit behavioural predictions.

The status of mental state attributions is heavily dependent on which account of social cognition one favours. Traditionally minded theory-theorists will be happy to say that the mental states attributed by our folk theory are essentially the same as those attributed by our social cognitive mechanisms, whilst simulation theorists and more liberal theory-theorists might draw a distinction between the kinds of states that we attribute explicitly and the kinds of states that are attributed by our implicit social cognitive mechanisms. Some, such as the interaction theorists, might even deny that our social cognitive mechanisms attribute any mental states at all. One benefit of my proposed distinction between social cognitive mechanisms and folk psychological discourse is that one can make sense of this denial whilst at the same time accepting that we engage in at least some explicit attribution of mental states and behaviour.

Narrative competency is typically understood as a capacity that we only draw upon when trying to make sense of complex social situations in which predictions of future behaviour are not immediately obvious. Likewise, normative constraints are usually understood as a product of our explicit folk discourse, although we may not always be explicitly aware of the influence that they exert on us. Nonetheless, one might wish to draw on both processes as part of an explanation of how our implicit social cognitive mechanisms are able to function. Ultimately where one draws the line between social cognitive mechanisms and folk psychological discourse is going to be heavily dependent on one's other theoretical commitments, and as we will see by the end of this thesis, keeping the two completely distinct is neither necessary nor desirable.

The most important point for my purposes is that we should keep the two kinds of phenomena at least in principle distinct from one another, even if in practice we discover that they overlap significantly. This 'in principle' distinction is necessary in order allow for the possibility that the content of our explicit folk psychological discourse is entirely unrelated to the content or structure of whatever implicit mechanisms implement social competency. This could be the case, for instance, if folk psychological discourse were nothing more than an exercise in rational reconstruction, where we create *post hoc* explanations for each other's (and our own) behaviours that bear no relation to the actual causes of those behaviours. Note that this could be true *even if* the underlying social cognitive mechanisms were able to accurately track the actual causes of behaviour – perhaps not at all likely, but nonetheless a coherent enough possibility that it is worth bearing in mind as a limiting edge case. Going forward it is important to keep this in principle distinction in mind, as a sort of foundation upon which we can begin our investigations of folk psychology and cognitive ontology.

This chapter has introduced a distinction between implicit social cognitive mechanisms on the one hand and explicit folk psychological discourse on the other. In chapters 2 and 3 I will look at how folk psychological discourse differs cross-culturally, and how it can serve a regulative as well as a predictive or explanatory role. In chapter 4 I will turn to cognitive scientific discourse, and consider how and when the explanations given by cognitive science differ from those given by folk psychology. I will also identify some historical cases where folk psychological intuitions and terminology have influenced cognitive scientific discourse. Finally, chapters 5 and 6 will take a more systematic look at the role played by folk psychological taxonomies in cognitive scientific explanations, and then propose a new methodology for replacing these taxonomies with more empirically supported terminology.

Chapter 2 – The Myth of a Universal Folk Psychology

This chapter will focus on cases of potential variation in both folk psychological discourse and social cognitive mechanisms. The focus will be primarily on the former, for two reasons. Firstly, there is more evidence of variation in folk psychological discourse than in social cognitive mechanisms. Secondly, the assumption that folk psychological discourse is universal is used implicitly as a way of licensing philosophers (and to some extent psychologists) to infer that the mental states posited by folk psychology are those that really exist. The aim in this chapter is to motivate one initial reason to be cautious of the role that folk psychological intuitions play in philosophy of mind and scientific psychology. By itself the argument presented is not sufficient to prove conclusively that folk psychology is unreliable, but it should at least give us cause to reconsider our reliance on folk psychological intuitions, and to explore alternative ways of categorising mental states.

In section 2.1 I will introduce what I call the universality assumption, and disambiguate two distinct versions of this assumption, along with the implications that they have for philosophy and cognitive science. In section 2.2 I will review studies of cultural variation in both social cognitive mechanisms and folk psychological discourse, and conclude that there is evidence of significant variation in the latter but not the former. In section 2.3 I will consider various critical responses to evidence of this kind. Finally, in section 2.4 I will argue that if we take the evidence of variation in folk psychological discourse seriously, we should be cautious about how we apply our own folk psychological intuitions.

2.1 – The Universality Assumption

In this section I will introduce two versions of the assumption that folk psychology is universal, one based on sub-personal social cognitive mechanisms and the other based on explicit folk psychological discourse. The former is *prima facie* more plausible than the latter, although examples will be given of theorists who endorse both claims. It is somewhat hard to find explicit endorsement of the universality

assumption (aside from by Fodor), precisely *because* it is an unspoken commonplace that no one usually thinks to mention. As such, this section will involve some reading between the lines in order to establish that the assumption does actually play an important theoretical role. I will argue that the main way that the assumption manifests itself is via what I call the 'no folk psychological miracles' argument: assuming that folk psychology is universal, and that it tracks real mental states, licenses theorists to simply take it for granted that the mental states posited by folk psychology are those that really exist.

2.1.1 – Levels of Explanation

Before going any further it will be useful to clarify some of the terms used when discussing 'levels of explanation'; terms such as personal/sub-personal, doxastic/subdoxastic, conscious/unconscious, and explicit/implicit. In this chapter I will be focusing primarily on a distinction between two kinds of folk psychological universality, which I will characterise as sub-personal and explicit respectively, but I will also refer back to the broader clarifications made here in future chapters.

Drayson (2012, 2014) draws attention to some philosophical confusion surrounding the use of the personal/sub-personal distinction. She distinguishes references to personal/sub-personal *explanations* on the one hand, and references to personal/sub-personal *states* on the other. The former is a distinction between two kinds of psychological explanation: personal level explanations are horizontal,[9] citing a sequence of events that led up to and caused the behaviour in question (Drayson 2012: 2-3), whilst sub-personal explanations are vertical, decomposing the person into component parts that together produce the behaviour to be explained. Drayson emphasises the fact that sub-personal explanations still typically rely on 'psychological' (as opposed to physiological) predicates, such as belief and desire –

[9] Drayson asserts that folk psychological explanations are typically horizontal, but here she has in mind the traditional interpretation of folk psychology as propositional attitude psychology. Under a broader definition of folk psychology, such as that which I argued for in the previous chapter, folk psychological explanations need not necessarily be horizontal.

what is distinctive about sub-personal explanation is that it applies such predicates to decomposed parts of persons rather than whole persons.

Personal/sub-personal states, meanwhile, are somewhat more complicated. A personal state is just a state of a whole person, such as believing that the sky is blue, and importantly does not necessarily need to be localised to any particular internal state (Drayson 2012: 9) – this is all perfectly consistent with how I argued folk psychology should be conceived in the previous chapter. Subsequently a sub-personal state is just a state of a sub-person, i.e. a belief attributed to a component part of a person, which only makes sense if one is giving a sub-personal explanation. For this reason Drayson argues that there is not really any distinction between personal and sub-personal states – they are exactly the same kind of thing, just applied in different ways or under different explanatory frameworks (*ibid*: 11). Importantly for Drayson, the kinds of states attributed in both personal and sub-personal explanation typically remain both intentional and distinctly psychological – whilst we can use non-psychological predicates in sub-personal explanations, this is not a defining feature of the sub-personal level.

To capture this further complication Drayson reintroduces the distinction between doxastic and sub-doxastic states (2012: 12), first described by Stich (1978). Doxastic states are simply those that correspond to standard folk psychological mental states, whilst subdoxastic states "don't correspond to anything posited by personal explanation" (Drayson 2012: 12). Both kinds of state can be used in sub-personal explanations, but the contents of doxastic states are potentially accessible to conscious introspection, whilst the contents of subdoxastic states are not. So beliefs, desires etc. attributed to sub-persons constitute doxastic states, whilst the attributions that figure in explanations of grammatical knowledge and early visual processing (for example) pick out subdoxastic states.

Drayson further clarifies that distinctions between the conscious and the unconscious, and between the normative and the non-normative, are also different to the personal and the sub-personal. What she calls the "cognitive unconscious" (2012: 15), i.e. the unconscious as studied by contemporary consciousness science, seems to

correspond to subdoxastic states, whilst the more traditional Freudian notion of the unconscious appears to be situated entirely at the personal level of explanation. The normative/non-normative distinction may pick out something interesting about personal level explanations, but not without further argumentation – I will return to this topic in the next chapter.

With all this in mind, I can clarify the distinction that I made in the previous chapter between social cognitive mechanisms and folk psychological discourse. The positing of social cognitive mechanisms is a form of sub-personal explanation, involving either doxastic or subdoxastic states. For example, the theory of mind posited by the theory-theory is (typically) a doxastic theory at the sub-personal level,[10] whilst the mental simulations posited by simulation theory are subdoxastic simulations, also at the sub-personal level. Folk psychological discourse, on the other hand, involves doxastic states at the personal level of explanation. So when I describe somebody as (explicitly or verbally) attributing belief or desire, or as (explicitly or verbally) situating someone else's behaviour in an on-going narrative, I am describing doxastic states at the personal level of explanation. I could also give a sub-personal level explanation of how that person is able to make those attributions, in terms of social cognitive mechanisms. The upshot of this distinction is that we could plausibly have universality at the level of social cognitive mechanisms *without* having universality at the level of folk psychological discourse, or *vice versa*. In the following subsections I will consider the evidence for and against both possibilities.

2.1.2 – Social Cognitive Universality

One way in which folk psychology could be universal is if the sub-personal mechanisms responsible for implementing it, i.e. the mechanisms studied by social cognition, were universal. We can call this **social cognitive universality**. Social cognitive universality would obtain if, for example, theory-theory were true and the

[10] Depending on how you spell out the details, theory-theory might be either non-doxastic or non-sub-personal, but I will not worry about that here. At least as it is most commonly interpreted, theory-theory seems to be a theory of doxastic states (such as beliefs and desire) implemented at the sub-personal level.

exact same (sub-personal) theory of mind was acquired in all cultures, or if simulation theory were true and all cultures performed (sub-personal) simulations in the exact same way. This could be the case even if there was variation in the verbal or behavioural expressions resulting from these sub-personal mechanisms. For example, cultural factors might prevent people talking about or acting on their knowledge of other minds, or might lead people to interpret attributions of belief very differently at the personal level. I will discuss this possibility in more detail in the next section.

There are a number of *prima facie* reasons for thinking that social cognitive universality might at least partially obtain. As soon as one concedes that there might be an evolutionary component to our explanations of social cognition, it seems plausible that at least the basic mechanisms will be shared across cultures (see e.g. Barkow, Cosmides & Tooby 1992; Carruthers, Laurence & Stich 2005). Of course, innateness is a distinct issue from universality, but it would be extremely unusual to find an evolved mechanism that was not at least somewhat universal (even if its expression depended on environmental factors). Even putting the evolutionary argument to one side, we might expect social cognitive mechanisms to converge on similar strategies for predicting behaviours that are shared across cultures, such as those motivated by innate drives like hunger, fear, and procreation. To the extent that one thinks human behaviour is universal, one might also expect social cognition to be universal.

However, human behaviour is of course not entirely universal, and neither is human cognition. Successful social cognitive mechanisms should be sensitive to cultural variations in behaviour, although this might not require the mechanisms themselves to vary (by analogy, the visual system could be both universal and capable of adjusting to novel environments). A more interesting result would be if the mechanisms themselves varied between cultures – say, hypothetically, if one culture relied more on simulation whilst another relied more on theoretical inference. In 2.2 I will consider evidence for this kind of variation and conclude that we in fact

have good reason to think that sub-personal social cognitive mechanisms are fairly universal.

2.1.3 – Folk Psychological Universality

Another way that folk psychology could be universal is if all cultures made the same kind of explicit claims about how other people think and behave. We can call this **folk psychological universality**, to distinguish it from the social cognitive universality that I described in the previous section. Whilst there is some overlap between this kind of universality and that which I described in the previous section (social cognitive universality), it also seems plausible that we might find more variation in explicit folk psychological reports, even if the underlying mechanisms were the same.

To illustrate this point, consider the following example: two groups of people have the same underlying social cognitive mechanisms, but differ in the way that they report their experience of other minds. In both cases the groups are able to systematically identify and categorise the same set of facial expressions, once that we might identify with the term 'anger'. However, the first group understands these facial expressions to be correlated with a kind of madness, and behaves accordingly, locking the person up but not thinking ill of them in anyway – they are a sick person to be treated, not just someone who cannot control their anger. The second group believes that controlling one's emotions is a moral imperative, and socially ostracises anyone who displays anger in this way. Both groups appear to detect the same 'mental state', but report it and respond to it in different ways that will have distinct cultural, social, and scientific implications. This is a fictional example, but I will discuss some real-world cases that might raise similar issues in the next section.

Nonetheless, even if it is plausible that there might be cultural variation in explicit folk psychological reports, the conventional assumption in philosophy has been that folk psychology will in fact turn out to be universal. Both Fodor and the Churchlands assume that the folk psychological theory over which they are arguing will be the same in all cultures, and furthermore that it will take the form of

propositional attitude psychology in particular. Fodor goes so far as to state "if an anthropologist claimed to have found such a group [that does not attribute beliefs and desires], I wouldn't believe him" (1987: 132). The Churchlands are less explicit, but it is apparent in their discussion of eliminative materialism that what they have in mind when they talk about the failure of folk psychology is a traditional propositional attitude psychology that is assumed to be universal across all cultures. When they criticise the folk theory for remaining static over several millennia, they have in mind a single theory, not a multitude of fragmented, culturally diverse theories. This is not to say that their criticisms are not valid, but rather that their choice of target reveals an inherent assumption of a universal propositional attitude psychology.

Stich does briefly consider the possibility that folk psychology might not be universal (1983: 217), quoting a passage where Hacking notes "ethnographers often find themselves quite unable to locate familiar mental states in alien cultures" (*ibid*; Hacking 1982: 44). Whilst acknowledging that this is an issue, he glosses over it swiftly by suggesting that the folk terms could simply be redefined so as to better suit the philosophical project in question. I will return to this issue in the next section after considering the evidence for cultural variation in explicit folk psychological discourse.

2.1.4 – No Folk Psychological Miracles

Whether or not social cognitive mechanisms are universal is an interesting question for the study of social cognition, and whether or not folk psychological discourse is universal is an interesting question for psychological anthropology. Both questions might be philosophically interesting insofar as one has a philosophical interest in social cognition and folk psychology, but what makes these questions especially interesting (for my purposes) is a further inferential step that I will call the 'no folk psychological miracles' argument. In brief, the argument is that *if* folk psychology were universal, then the best way of explaining this universality would be that it accurately captures how minds actually work, and therefore licenses us to use our

folk psychological intuitions as a guide to cognitive scientific discovery. Some version of this argument seems to be implicit in the work of those who take folk psychology to be universal in some sense, and in this subsection I will formulate the argument explicitly and consider its implications.

In philosophy of science the 'no miracles' argument originates with Putnam's claim that realism "is the only philosophy that doesn't make the success of science a miracle" (1975: 73). The thought here is that if a successful scientific theory posits the existence of certain unobservable entities, then the best way of explaining the success of that theory is to accept that those entities really do exist. For example, our best theories in physics and chemistry seem to be committed to the existence of electrons, even though we cannot directly observe them. If electrons did not exist than the success of these theories would appear to be miraculous; therefore, electrons exist. Whilst the status of this argument is hotly contested in philosophy of science (see Chakravarrty 2015, especially sections 2.1, 3.2, and 3.3), it seems to be a prima facie plausible way of reasoning about unobservable entities, and something very much like it seems to underlie the inference from folk psychological universality to the use of folk psychological concepts in cognitive science.

By analogy, if folk psychological discourse is universally committed to the existence of the same kinds of mental states and processes, and enjoys explanatory and predictive success, then we might feel licensed to make the inference that these mental states and processes actually do exist – otherwise, the success of folk psychology would be a miracle. Fodor seems to be implicitly committed to this kind of argument when he populates his language of thought with propositional attitudes drawn from his own folk psychological discourse, as do many other philosophers who simply assume that the correct psychological theory is that posited by folk psychology (and that there is only one such theory). In an odd way the Churchlands' are also committed to at least the logic of this argument, in that they target the *success* (or lack thereof) of folk psychology as a reason for or against elimination, rather than simply denying that we should be basing our scientific psychology on our folk psychological intuitions in the first place. In contrast, my argument will be that

even if folk psychological discourse exhibits explanatory and predictive success, there are still reasons that we should not use it as the basis for our scientific theories of the mind.

2.2 – Evidence For and Against Universality

In the previous section I characterised two versions of the universality assumption, one claiming that sub-personal social cognitive mechanisms are universal, and the other claiming that explicit folk psychological discourse is universal. For most of the 20th century these two assumptions went unquestioned, but in the last couple of decades they have come under scrutiny from a number of distinct angles. Below I will introduce some initial reasons for being sceptical about the universality of folk psychology. In the rest of the section I will consider evidence of variation from four different fields: social cognition, anthropology, comparative linguistics, and experimental philosophy.

Lillard (1998) identifies "an assumption that everyday, unschooled knowledge of human psychology is basically the same everywhere" (*ibid*: 3), and suggested that if this assumption were false it might call into question some central findings in social cognition. She provides a useful definition of this assumed universal knowledge of human psychology as "the naïve folk psychology described or implied by the late 20th century academic literature on psychology and philosophy of mind" (*ibid.*), which she refers to as the 'European-American Social Sciences Model' (EASSM). This is essentially equivalent to the folk psychological discourse that philosophers such as Fodor and the Churchlands assume to be universal. Lillard also notes an additional subtlety, which is that there may be variation of this model *within* European and American cultures, as thus far most research in social cognition has primarily targeted a relatively small population centered in and around university departments.

Henrich, Heine, & Norenzayan make a similar point in their now-classic paper "The weirdest people in the world?" (2010), where they introduce the acronym WEIRD (Western, Educated, Industrialized, Rich, and Democratic) to refer to the

societies that have provided the subjects for the majority of experiments for pretty much the entire history of experimental psychology. Their concern is that as "there are no obvious a priori grounds for claiming that a particular behavioral phenomenon is universal based on sampling from a single subpopulation" (*ibid.*), it is possible that psychological science may be systematically biased by the limited (WEIRD) population it draws its samples from. My focus in this section is on challenging the assumption that the folk psychological discourse common to WEIRD/EASSM societies will generalize unproblematically to humankind as a whole.

When looking at the evidence either for or against the universality assumption, it is important to keep in mind the distinction that I introduced in the previous section. There may be evidence for cultural variation in what people explicitly say about other minds without there being evidence for variation in the sub-personal mechanisms responsible for social cognition. Astuti makes this point very clear when she writes,

> we should not be tempted into using folk theories of the mind […] to make claims about cross-cultural variation in people's Theory of Mind. This is because people's folk theories are explicit reflections about the human mind, whereas Theory of Mind – as understood by psychologists – operates implicitly. (Astuti 2014)

I will first consider evidence for and against the assumption that sub-personal social cognitive mechanisms are universal, and conclude that this assumption is probably a fairly reasonable one. I will then look at several kinds of evidence for and against the assumption that explicit folk psychological discourse is universal, and conclude that this is a questionable assumption with potentially serious implications for theoretical work in philosophy and cognitive science.

2.2.1 – Social Cognition

As I discussed in the previous chapter, the classic paradigm in social cognition is the false belief task, which is used as a proxy for general social cognitive development. The standard versions of the task, where infants are asked where they think the actor

or doll will look, are referred to in the literature as 'elicited response' tasks. Whilst the elicited response version of the false belief task does typically involve an explicit verbal report, it is intended to track the capacity of sub-personal social cognitive mechanisms, and as such falls under the category of the first kind of potential universality that I discussed above. Nonverbal tasks, such as those performed by Onishi & Baillargeon (2005), may be more directly related to processing at the sub-personal (and perhaps subdoxastic) level, as they do not require an explicit verbal report.

WEIRD children are typically able to pass the elicited response task around the age of four or five, and some versions of the nonverbal task significantly earlier, perhaps even as early as 13 months (Surian, Caldi, & Sperber 2007) although the evidence for this is controversial. Originally this result was taken as evidence for a standard developmental trajectory, with autistic childrens' slow development providing an interesting contrast, but more recently there have been some attempts at conducting false belief tasks in other cultures and contexts as a way of confirming whether or not the development of this capacity does in fact follow a universal pattern.

One of the earliest attempts to explore cross-cultural variation in false belief acquisition was carried out by Vinden (1996), who conducted a culturally appropriate version of the elicited response task with Junín Quechua children living in the Peruvian Andes. The Junín Quechua culture is traditionally oral, with a language that refers only indirectly to mental state concepts, e.g. by using a word more like "say" where other cultures might use "think". The children performed relatively poorly on the false belief task, answering no better than chance even up to the age of 7 (the oldest included in the study). Vinden concluded that Junín Quechua children develop a theory of mind competency later than comparable "Western literate children".

In the two decades since Vinden's study there has been a more concerted effort to explore potential variation in false belief acquisition, with the general consensus being that there is a small degree of variation in developmental trajectory

across cultures, but no major variation in eventual outcome (i.e. all neurotypical children do eventually acquire an understanding of false beliefs). Wellman *et al* (2001) found variation in timing of false belief acquisition across several cultures, but had a relatively small sample size in non-European American cultures. Callaghan *et al* (2005), in contrast, found no significant variation across Canadian, Indian, Peruvian, Samoan and Thai children. Naito & Koyama (2006) focused specifically on Japanese children, and found a delay of roughly one year in acquisition of false belief understanding, which they attributed to a greater emphasis on environmental causes of behaviour in Japanese culture (*ibid*: 300). Finally, Liu *et al* (2008) found parallel development in Chinese and North American children, in contrast to greater internal (primarily socio-economic) variation in both cultures. It is important to note that in all of the above cases the tests conducted used the elicited response paradigm, and it appears that there is significantly less variation when non-verbal paradigms are used (see e.g. Barrett *et al* 2013), which are plausibly less vulnerable to cultural or linguistic interference. Lavelle (2016) presents further discussion of this data, and argues that there are a number of distinct ways in which the developmental delay in explicit false belief understanding found in some cultures can be accounted for. Nonetheless, it does seem to be true that children across all cultures do eventually pass the false belief task, and even if culture does have an impact on social cognitive mechanisms, it does not seem to be an especially pronounced one. In the rest of this chapter I will focus primarily on variation in explicit folk psychological discourse.

2.2.2 – Anthropology

The primary source of evidence against folk psychological universality comes from the anthropological study of other cultures' folk theories of minds, sometimes referred to as 'ethnopsychology' (see e.g. Howard 1985, White 1992). Lillard (1998) provides an overview of such evidence, broken up into four main categories (which she notes are not necessarily mutually exclusive):

- *Attraction to Magic* – some cultures allow for the mind to interact with the world in ways which scientific psychology would deem 'supernatural', e.g. extrasensory perception or spiritual influence.
- *Differing Conceptual Distinctions* – some cultures categorise aspects of cognition in different ways, e.g. distinguishing senses or attitudes that other cultures might not recognise.
- *Denial of the Negative* – some cultures refuse to openly acknowledge negative emotions or feelings, e.g. anger or sadness.
- *Different Values* – some cultures place differing emphasis on aspects of cognition, e.g. internal states vs. external behaviour.

  (Adapted from Lillard 1998: 23-4)

I am most interested in the second (and to some extent the fourth) of these categories. Unless we are to contemplate a radical shift in the naturalistic understanding of cognition, the first must remain a mere curiosity (at least so far as scientific psychology is concerned). The third does not seem to speak to anything particularly fundamental – whilst some cultures might refuse to speak of certain mental states, presumably they still experience something akin to anger or sadness (recall our own Victorian ancestors' attitude towards open displays of emotion).

Differences of emphasis, in particular between internal/external factors, could be seen as contributing to some extent to the current dominant approach in cognitive science. What Lillard calls the "European-American" model places heavy emphasis on internal states, demonstrated by the "large and varied vocabulary [that] EAs use to refer to emotions and other mental processes". In contrast, other cultures have a far more limited vocabulary in this area, and in some cases do not even talk about the mind, let alone its content (see Lillard 1998: 12-13). Whilst this might simply reflect the influence that post-Cartesian philosophy and psychology has had on European-American culture, it is not inconceivable that this culture has in turn influenced scientific developments, leading experimentation and theorising towards what might

broadly be considered a "representational" (or propositional) paradigm that focuses on the interaction of internal states rather than a more general description of behaviour.

An extreme version of this can be found in the so-called "opacity of mind" doctrine attributed to some Pacific island cultures (see Robbins & Rumsey 2008). These cultures are reported to make no explicit reference to the mental states of others, which is regarded as a taboo subject. Whilst the practical implications of this have been somewhat overstated in the past, and whilst the people from these cultures undoubtedly have some awareness that other people may have internal mental lives, it is still interesting to imagine what a philosophy of mind (or perhaps we should say a philosophy of *behaviour*) could look like under such circumstances.

Of especial interest are the divergent conceptual distinctions that we find across different cultures. If we accept that our study of the mind and brain is likely going to be shaped by the concepts available to us, then it makes sense to pay attention to potential alternative conceptual schemes. This is not to say that these alternatives will necessarily be superior, but at the very least they will highlight the contingent nature of our own concepts. I consider some examples of this based on lexical variance in the next section, but we can also identify conceptual distinctions that are embodied in broader cultural practices, such as whether a distinction between intentional and unintentional action is recognised or given moral weight (cf. Barrett *et al* 2016).

More recently, a collection of short position papers from an interdisciplinary conference titled "Toward An Anthropological Theory of Mind" (Luhrmann 2011) gives a flavour of the kinds of cultural variation that can be found in folk psychological discourse. A particularly interesting example is the observation that "the philosophically important 'believe' (*tz'ok-es*) is only fully appropriate in Mopan [a Yucatan Mayan language] usage if the 'believer' also 'obeys' what s/he is told" (Danziger 2011: 52). Interpreted from within the WEIRD folk psychological framework, this would seem to reflect a greater level of commitment than we usually ascribe to belief. At least according to the standard philosophical account, there is

nothing inconsistent about acting contrary to one's beliefs, for example if you believe that something is morally wrong but do it anyway. *Tz'ok-es* is perhaps closer to what we mean by religious belief, i.e. a belief that strongly commits you to certain future behaviours, and which no longer applies to you if you stop behaving in those ways. It certainly has a distinct functional profile, and direct translation between *believe* and *tz'ok-es* could lead to terminological confusion. Similarly, Matthews (2013) notes that there is "considerable evidence of cross-cultural variation in the role played by propositional attitudes in commonsense psychological explanations" (*ibid*: 110), citing Vinden (1996) in support of his claim that "central Peru's Junín Quechuan culture [….] reportedly makes little or no use of propositional attitude attributions" (*ibid*). He goes on to discuss how such cultures seem to describe behaviour in terms of contextual factors, rather than by attributing mental states.

What evidence like this suggests is that the explicit, personal level content of folk psychological discourse might not be quite so universal as has been previously assumed. Even if states similar to the propositional attitudes of WEIRD folk psychology can be identified, they might be emphasised or applied in sufficiently different ways to make straightforward one-to-one translation problematic. At the very least it raises the possibility that our own folk psychological taxonomy is just as liable to subtle differences in emphasis, which should lead us to be wary of using it as the sole basis for our scientific taxonomy.

2.2.3 – Comparative Linguistics

Another interesting source of evidence for folk psychological variation comes from studies of linguistic variation in concepts referring to the mind. Although such evidence is by no means conclusive, it can give us at least an initial indication of alternative folk psychological conceptions of the mind. The Hausa of Nigeria only distinguish (lexically) between two sensory channels: *gani* (sight) and *ji* (hearing, tasting, smelling, touching, intuition, and knowing), rather than the five senses typically identified in the European-American folk taxonomy (Ritchie 1991; cf. Lillard 1998: 19). Whilst it is contextually apparent which sense they are referring to,

they clearly place less emphasis on the distinctions between the senses other than sight.

In Japanese there is no simple distinction between the mental and the non-mental, as is normally assumed in philosophical interpretations of folk psychology, but rather a number of different concepts including *kokoro, hara, ki, seishin,* and *mi*, none of which clearly refer either to the mind or the body alone. Lillard summarises these different concepts:

> *Kokoro*, sometimes translated as "heart, feeling, spirit, intention, will, mind" is best translated as "the embodied mind" (p. 63), according to Lebra, in part because it has a strong emotional component that is usually not considered part of the more rationalistic EA [European-American] mind concept. For EAs, minds interpret events and thereby give rise to emotions, but their primary force is in cognition. Rather than being placed within a thinking head, *kokoro* is located in the heart and has strong links to blood and genes. Moving along a continuum from *kokoro* toward ethereal or spiritual selves are the terms *hara*, "the vital center of the body-mind"; "inner state" or *ki*, which "circulates throughout a person's body-mind" (p. 64); and *seishin*, which is even more closely linked to spirit. At the other end of the spectrum, *mi* refers to the body, but it is a body permeated with mind, combining "spirit and body, mentation and sensation, the conscious and unconscious . . . not a fixed entity but a 'relational unity' which emerges out of involvement with other (persons or things)" (p. 65). This is clearly different form the EASSM of mind, not simply a difference in emphasis. These distinctions fit into an entirely different conceptual landscape. (Lillard 1998: 12; references to Lebra 1993)

Whilst it could be argued that these novel distinctions reflect cultural idiosyncrasies rather than any fundamental variation in folk psychological discourse, the very fact that these divergent conceptual taxonomies are possible should at least lead us to question the necessity of our own folk psychological distinctions. If contemporary analytic philosophy of mind had developed in a culture with the conceptual distinctions present in Japanese language, would it be so focused on distinguishing the mental from the non-mental, or would it even make sense to talk of a mind-body problem?

Focusing on a more specific folk psychological domain, Wierzbicka has argued that a language's emotion terms "constitute a folk taxonomy, not an objective, culture-free analytical framework" (1986: 584), and as such we should not base our scientific study of emotions on categories drawn from natural language. She gives two illustrative examples:

> Polish does not have a word corresponding exactly to the English word *disgust*. What if the psychologists working on the "fundamental human emotions" happened to be native speakers of Polish rather than English? Would it still have occurred to them to include "disgust" on their list? And Australian Aboriginal language Gidjingali does not seem to distinguish lexically "fear" from "shame," subsuming feelings kindred to those identified by the English words *fear* and *shame* under one lexical item (Hiatt 1978: 185). If the researchers happened to be native speakers of Gidjingali rather than English, would it still have occurred to them to claim that fear and shame are both fundamental human emotions, discrete and clearly separated from each other? (Wierzbicka 1986: 584)

I will discuss Wierzbicka's proposed solution, the creation of a "natural semantic metalanguage", in my final chapter on cognitive ontology revision, but I introduce her work here in order to give an indication of the kind of folk psychological variation that I think problematizes the universality assumption. If our philosophical and scientific taxonomies are artificially limited by the concepts available in our language, then the apparent falsehood of the universality assumption (at least with regard to explicit folk psychological discourse) could have serious theoretical implications.[11]

[11] I am aware that all of the evidence discussed in this section raises the spectre of the now controversial Sapir-Whorf hypothesis. Whilst I have no particular sympathy for the strongest versions of this hypothesis, there has been some interesting progress made towards rehabilitating weaker and more plausible versions of linguistic relativity (see e.g. Regier & Kay 2009), and I do find it plausible that linguistic variations in folk psychological discourse might at least be indicative of distinct cultural schemas for understanding other minds. This is not to say that the members of these cultures perceive the world differently, or that this linguistic variation has a strong impact on basic social cognitive mechanisms, but rather that when it comes to the explicit expression of how one understands other minds, it seems natural to think that this expression might be mediated through a socio-cultural lens of which language is just one small part. Whether these linguistic differences are in fact the cause of the cultural variation, or whether they are simply caused by it, seems to me to be a moot point that ignores

2.2.4 – Experimental Philosophy

A more recent source of evidence against the universality assumption comes from so-called 'experimental philosophy', which aims to test philosophical intuitions across a wide range of diverse populations. In this section I will discuss some recent work in experimental philosophy that has focused on variation in folk psychological intuitions. Experimental philosophy is a somewhat controversial area of research – in the next section I will consider and then respond to some criticisms of this approach, which may also be relevant to my more general argument against the universality assumption.

Systma (2014) presents a collection of recent work on experimental philosophy of mind, which he describes as being in the business of testing people's intuitions about attributions of phenomenal states, "like feeling pains, seeing colours, hearings sounds, and so on" (*ibid*: 3), rather than attributions of mental states such as belief and desire. Whilst this is a somewhat idiosyncratic definition, as regular philosophy of mind is certainly concerned with both phenomenal and non-phenomenal mental states, it does help distinguish experimental philosophy of mind from other areas of experimental philosophy on the one hand, and cross-cultural psychology or social cognition on the other. I will therefore focus on intuitions about attributions of phenomenal states in this section, although I see no reason why experimental philosophy of mind should not also concern itself with attributions of non-phenomenal mental states.

Two of the papers in Systma's collection focus on Block's (1978) Chinese nation thought experiment, which asks us to imagine that the entire population of China have been connected together in such a way as to approximate the functional organisation of a human brain, and then questions whether China would thus be endowed with mental states. Block's intuition is that it would not, and therefore that functionalism is false, but Nado (2014) notes that this intuition may not be universal.

---

the complex and recursive interactions between language, culture, and cognition. It is entirely possible that language is at once shaped by cultural influences and also involved in shaping the culture that it is part of.

Knobe & Prinz (2008) found that whilst their subjects intuitively hesitated to attribute *phenomenal* states to group entities (such as the Chinese nation), they were more comfortable attributing non-phenomenal mental states to groups, and as such might be willing to grant that the Chinese nation can think, but not that it can feel. Huebner *et al* (2010) found that subjects in Hong Kong were less hesitant to ascribe phenomenal states to groups, so there may also be cultural variation in intuitions about Chinese nation style cases. If this is the case then it may be problematic for Block to base a philosophical argument upon the strength of his own intuitions alone (although see below for possible responses to this line of argument).

Buckwalter & Phelan (2014) present evidence that folk intuitions generally allow that disembodied agents (such as ghosts and spirits) can undergo phenomenal experience; supposedly disproving a common philosophical claim that having the right sort of body is a necessary condition for mentality (Block's intuition that the Chinese nation cannot have mental states seems to be a version of this claim). Once again, the fact that folk intuitions diverge from those of philosophers in these cases might lead us to question the role of intuition in philosophical argumentation. However in this case one could accuse the folk of engaging in the kind of magical thinking that Lillard discusses – given that our naturalistic philosophical ontology does not include ghosts and spirits, one could question whether we ought to respect intuitions about such entities.

The rest of the volume continues in much the same way. Reuter *et al* (2014) present evidence that the everyday concept of pain does allow for pain hallucinations, contrary to a common philosophical position that pain hallucinations are conceptually impossible. Tierney *et al* (2014) argue that folk intuitions support a pluralism about personal identity, and Machery (2014) demonstrates that native English speakers "are willing to endorse a surprising range of seemingly contradictory sentences" (Systma 2014: 8), which he uses to support his heterogeneity thesis about psychological concepts (see 5.3.1 for further discussion of Machery's concept eliminativism). The details of these papers are not especially important for my current purposes – what matters is that folk intuitions about the

mind seem to frequently differ from those that philosophers appeal to in their arguments, and furthermore often differ between populations or samples.[12] This being the case, the implicit commitment to folk psychological universality that I identified in 2.1 starts to look increasingly unsustainable, and rejecting this assumption may have serious implications for both philosophy and cognitive science. In the next section (2.3) I will consider how a proponent of the universality assumption might respond to this evidence, before finally discussing what the rejection of the universality assumption entails (in 2.4).

2.3 – Accounting For The Evidence

The evidence that I presented in the previous section indicates that there is at least some cultural variation in explicit folk psychological discourse, even if the sub-personal mechanisms responsible for social cognition may turn out to be universal. In the next section I will argue that this variation gives us good reason to be cautious of relying too heavily on our own folk psychological intuitions, but first I will consider some possible alternative responses.

Machery (forthcoming) presents a comprehensive overview of several common criticisms of experimental philosophy, some of which can be generalised to apply to the approach that I am taking in this chapter (Nado 2014 presents a similar list of responses). The first response is to simply criticise the methodology of experimental philosophers (Machery forthcoming: 235-44). In 2.3.1 I will consider one particular methodological issue that is especially relevant here, which is the problems that occur when attempting to translate technical terms. The second response is to claim that the intuitions of experts (such as philosophers) should be given more credence than those of the general public (*ibid*: 244-261) – I consider

[12] We should note that some of this variation is not cross-cultural in the geographic sense, but rather reflective of variation between different parts of the same (geographical) culture, i.e. variation between typically upper-middle class and well-educated philosophers on the one hand, and the rest of the 'folk' on the other. Nonetheless, as Lillard (1998) notes explicitly, this 'academic' culture from which most of our scientific and philosophical intuitions are drawn is certainly distinct enough that we should at least question the universal applicability of those intuitions (cf. Henrich, Heine & Norenzayan 2010).

how this applies to the universality assumption 2.3.2. Another interesting suggestion he raises is that we could reform our use of intuitions (*ibid*: 273-5) – I will consider something like this in 2.3.3 and 2.3.4. The remaining responses are all more specific to experimental philosophy, and to the method of cases in particular, but I will also consider some themes that emerge here, especially regarding the possibility that some kinds of intuitions might survive even if others prove to be unsuitable. The further issues Machery mentions are that results from experimental philosophy: might not generalise (*ibid*: 241-66); might reflect fallibility rather than unreliability (*ibid*: 266-73); might mischaracterise the role of intuitions (*ibid*: 275-8); and might overgeneralise (*ibid*: 278-84). Each of these may also apply to evidence of variation in folk psychology, but I will not respond to them directly here.

2.3.1 – Translation Errors

Problems can arise when we translate concepts in order to test their use in other cultures. We can end up drawing unwarranted conclusions based on a bad translation, although the very fact that translation is difficult might suggest that the concept in question is not entirely universal. Wierzbicka relates how, where a language lacks a term for a particularly complex emotion, the speakers of that language will sometimes come up with a more long-winded way of expressing that emotion (1986: 587). Similarly, even if a certain language lacked a term for some mental state that we typically take to be basic, it might turn out that speakers of that language have a more circuitous way of expressing the same concept. Simple transliteration of a language's folk psychological concepts will be liable to miss such subtlety. Lillard is also sympathetic to this issue, noting that despite the Hausa word *ji* referring to any sense other than sight, "one can generally tell from context whether something is smelled or tasted" (1998: 19). The linguistic variation marks a difference in emphasis rather than a full-blown conceptual shift. Indeed, given the physical similarity of people from different cultures (we all have the same basic sense organs) it would be extremely surprising to find a culture that could not distinguish smell and taste, even in the absence of a lexical distinction.

Nonetheless, even if it were the case that two terms could be translated so as to refer to the same concept (given a sufficiently refined translation process) the very fact that this translation process is non-trivial might give us reason to think that the universality assumption is unsustainable, or at least that the inference from universality thesis to acceptability of folk concepts is ill-advised. Consider that the difficulty of translating between certain folk psychological terms is partly due to the additional contextual information that is needed in order to make sense of how the term is being used in each particular case. So whilst within a certain context the term *ji* unambiguously refers to either smell or taste, it becomes ambiguous the instant one removes it from this context. And the scientific or philosophically application of these terms is precisely one where they are likely to have been stripped of context, rendering them unhelpfully vague or indeterminate. Consider a science of the senses that used the term *ji* – without further clarification, such as writing "ji (distal chemical)", it would be impossible to tell how this term was being used each time it was written. Note this is not just a problem with non-WEIRD languages – as we will see in chapter 4, there are many cases where the folk psychological terms used in European-American philosophy and cognitive science are just as vague. So whilst cases of translation error might seem to support a version of the universality assumption against charges of cross-cultural variation, they nonetheless highlight the potential for terminological and conceptual misunderstandings. In chapter 6 I will argue that the easiest way to avoid such misunderstandings is to adopt a novel, non-folk psychological taxonomy.

2.3.2 – Expert Intuitions

Hales (2006) has argued that the intuitions of the experts in any given field should be granted greater credence, and Williamson (2007) has explicitly claimed that philosophers' intuitions about thought experiments should be expected to be more reliable than those of the general public. This certainly seems reasonable for some disciplines – for example, you should probably trust a trained civil engineer's intuitions about the safety of a bridge, or a trained neurosurgeon's intuitions about

the weird blob on your brain scan. Similarly, trained philosophers may have a better grasp of philosophical concepts, and more experience of applying those concepts to thought experiments (Machery forthcoming: 246-7). If this were true then the evidence from experimental philosophy might merely be a demonstration of the training required in order to do good philosophy, just as evidence of variation in intuitions about bridge safety might just be a demonstration that most people don't know anything about bridges.

However, the analogy between scientific expertise and philosophical expertise does not seem entirely clear. An engineer's intuitions about the safety of a bridge can ultimately be verified by investigating the bridge itself, as can a neurosurgeon's intuitions about your brain. Philosopher's intuitions, on the other hand, cannot be independently verified, and may just reflect the accepted 'truths' of the tradition they have been educated within (cf. Weinberg *et al* 2010). Furthermore, there is evidence of substantial disagreement and biasing effects between philosophers (Nado 2014; Machery forthcoming: 255-61), which we would not expect in the case of scientific intuitions – indeed, if we found such disagreement or biasing we would probably discount those intuitions in favour of more substantial empirical investigation.

Even granting that the expertise argument may hold for some cases in experimental philosophy, it is not at all clear how it is meant to work for folk psychological intuitions. What I am interested in are not intuitions about abstract thought experiments in philosophy of mind, but rather everyday intuitions about real-world minds. Here the best candidates for experts are not the philosophers or the psychologists, but just the folk themselves. We are all (relatively) competent predictors of each other's behaviours, at least under normal conditions, and there is no evidence that philosophers or psychologists are any better in this regard. Indeed, it is precisely the everyday, 'manifest' image of behaviour that folk psychology is supposed to describe, not just a particular theory of behaviour that is restricted to the educated intelligentsia. Of course, when a philosopher or psychologist applies themselves to technical issues in cognitive science they may do better than a random

member of the public, but this is distinct from saying that this expertise crosses over into their everyday intuitions about other people's minds. It seems fair to say that we all enjoy an equal level of expertise when it comes to everyday folk psychology (or at least, that philosopher and psychologists are certainly no better, on average, than anyone else).

2.3.3 – Conceptual Convergence

Even if folk psychological intuitions appear to exhibit cultural variation, we might be able to explain this away as either an interpretive failure or an irrelevant cultural gloss. Stich suggests something along these lines when he considers the possibility that philosophy might be focusing too heavily on belief and desire, potentially at the expense of other culture's folk theories of mind (1983: 217). He suggests that what many theorists do at this point is simply to redefine 'belief' and 'desire' such that they cover any potential propositional attitude, but notes that this may just distort the terms so far that they cease to bear any resemblance to their folk usage (*ibid*: 218). A similar strategy is implicit in Lewis' proposal that we can extract a folk theory of mind by operationalizing "folk psychological platitudes" (Lewis 1972; see 1.1.3) – presumably this operationalization could include a function that averages out cultural variation? In chapter 6 I will consider the possibility of using cultural *universals* as the basis for a revised cognitive ontology, drawing on a suggestion made by Turner (2012), but for now I will just note that a strategy of this kind might be able to account for at least some of the apparent variation.

A related group of responses to the evidence from cross-cultural research suggests that even if some intuitions (i.e. those directly targeted by the research) are problematic, we should not apply our scepticism to intuitions in general. Alternatively, another version of this response says that we could try to improve the process of eliciting intuitions so as to make it more reliable, perhaps by standardising the context and manner in which thought experiments are presented. However, the problem with this suggestion is that we have no real sense of what would constitute a 'standard' format for thought experiments – perhaps the way that thought

experiments are presented in undergraduate tutorials is actually very unusual and elicits strange intuitions? The problem is even worse when it comes to folk psychological intuitions, which are surely most valid when they occur *within* their original cultural context.

2.3.4 – The Disambiguation Strategy

One common response to evidence of variation in intuitions is to argue that the populations whose intuitions vary might just differ in their understanding of the questions being asked or the concepts being deployed. For instance, Sosa (2009) has argued that the apparent variations in Gettier case intuitions demonstrated by Weinberg *et al* (2001) might just be evidence of two distinct ways of interpreting the term 'knowledge', rather than actual variation in intuitions. The thought here is that if you were able to get both samples to agree on a single interpretation of 'knowledge', then their intuitions would turn out to be the same. A similar argument could be made concerning the variation in intuitions about attributions of phenomenal states to groups – perhaps people from Hong Kong just have a different understanding of what it means to be conscious than people from the UK? Sosa's suggestion is that we just embrace a terminological pluralism, and accept that there are different ways of using technical terms such as 'knowledge' and 'consciousness'.

I am actually very sympathetic to this suggestion, but I think it demonstrates an important consequence of acknowledging cultural variation, rather than a refutation of it. It might turn out that cultural variation in the application of philosophical concepts can reveal to us ways in which our own understanding of those concepts was limited. So perhaps it turns out that disambiguating consciousness so as to distinguish between group consciousness and individual consciousness is a useful technical innovation (see Irvine 2013 for other ways in which consciousness might be disambiguated). In chapter 4 I will explore this 'disambiguation strategy' in more detail, and apply it to several distinct case studies drawn from both philosophy and cognitive science. Cultural variation may provide an initial motivation for adopting this strategy, but I will argue that the strategy has

independent value as an approach to resolving the theoretical ambiguities that can occur when we adopt imprecise folk psychological concepts.

Another version of this strategy would be to argue that folk psychological variation might just be evidence of actual variation in cognitive processes. Variation in folk intuitions about how the mind works could reflect *actual* variation in how the minds works – we could all possess the exact same intuition forming processes, but if the subject of those intuitions, i.e. other minds, were different, then we would naturally reach different conclusions about those minds for epistemically sound reasons. I think this is a valid concern, and one that has potentially serious implications for our scientific study of the mind, but it is important to distinguish a couple of distinct ways in which it could play out. The first would be to claim that there is innate (i.e. genetically coded) variation between different human populations, and that this results in variation in how our minds work. Whilst it is possible that this could be the case, we do not yet have any clear understanding of the genetic basis of cognition, and given the relative genetic homogeneity of human populations it seems unlikely that this could account for all behavioural variation. The second, more plausible interpretation of this argument is that there are cultural and environmental factors that lead to variation in the expression of cognitive processes. For example, there is some evidence that visual illusions based on straight edges and corners (e.g. the Müller-Lyer illusion) only affect subjects from cultures where rectangular buildings are the norm (Segall *et al* 1966; Ahluwalia 1978), which might lead to divergent folk psychological intuitions based on genuinely divergent cognitive processing. Someone from one of these cultures would presumably make different predictions about the behaviour of subjects exposed to illusions of this kind. Another interesting way in which culture could affect the mind is via so-called 'mindshaping' mechanisms, where folk psychological discourse itself has a recursive effect on how we think and behave. Such mechanisms will be the topic of the next chapter, but for now I will simply say that either way this plays out, variation in folk psychological discourse will still have important implications for how we use folk psychological concepts in philosophy and cognitive science.

2.3.5 – Uncovering Hidden Universals

Both Wierzbicka and Lillard express a hope that despite all of the variation, some universal aspects of folk psychology might yet be uncovered. If this were possible then it might give us some clues as to the genuine structure of cognition – or alternatively, it might just indicate that there are some innate (or commonly occurring) folk psychological principles that are nonetheless wrong. Regardless of the latter possibility, finding any universals at all would surely put us in a better place (*qua* constructing an accurate scientific and philosophical lexicon) than simply continuing to follow our (culturally and linguistically biased) folk psychological instincts.

Lillard proposes looking for universals in three places: the literature on primate theory of mind, the developmental literature, and the ethnopsychological literature (1998: 26). If we can identify mindreading mechanisms that we share with our closest relatives, then presumably they evolved long enough ago to be common to all modern humans. Similarly, if an aspect of folk psychology develops before cultural influences can really set in, then we might have reason to think that it is universal. Finally, we can try and identify features common to all folk psychologies, or perhaps try and 'average out' the differences in order to find more-or-less universal features.

As Lillard recognises, each of these suggestions comes with some associated difficulties. Aside from the obvious practical issues that arise when working with animals, children, or cultures very different to ones own, there is an additional risk that any evidence of universality that is uncovered will be tainted by the researcher's pre-existing folk psychological intuitions. This is most apparent when working with other cultures, where an ethnographer's instincts and native language might shape the way that they interpret their subjects. Attempts to conduct cross-cultural versions of basic social cognition experiments often run into similar trouble. Povinelli & Vonk (2003) have even argued that comparative research is not able to escape the

touch of lexical bias, criticising the use of anthropocentric mental state attributions when describing the behaviour of non-human primates.

Lillard suggests a partial solution to these problems, that "researchers should give carefully constructed, culturally sensitive tests [of folk psychological intuitions]" (1998: 27). This would certainly help to give a clearer picture of the scale of folk psychological variation, but it would not overcome the basic problem identified by Wierzbicka, i.e. that the language we express our theories with is itself liable to shape those theories. It is also less obviously applicable to primate research, although better experimental design is certainly to be encouraged there as well.

Wierzbicka's 'Natural Semantic Metalanguage' (NSM) superficially resembles Lillard's third proposal, in the sense that it is an attempt to 'boil down' cultural differences in order to find a basic, universal core shared by all cultures. It is far more ambitious, however, aiming to apply not just to folk psychology, but also to natural language more generally. The strategy, Wierzbicka writes, "is based on the assumption that the shared core of human thought is reflected in the shared core of all languages and can be identified through empirical linguistic investigations" (2005: 259). Through such investigation she claims to have discovered the set of basic 'conceptual primes' that are shared by all languages and cultures.

I will not assess Wierzbicka's project here,[13] but regardless of its success or failure her approach is useful in that it draws attention to the conceptual contingencies of natural language, and attempts to provide a neutral platform on which to discuss such contingencies. Moving forward, we will need something like this in order to discuss the shortcomings of folk psychology in relation to contemporary scientific psychology.

There is another possibility that neither Wierzbicka nor Lillard consider. This is Turner's (2012) suggestion that we should construct an entirely new conceptual lexicon, drawing not only on ethnographic and linguistic research but also on data from neuroscience and cognitive psychology. The aim of such a lexicon would be to

[13] Although see Drobnak (2009) for some criticisms of her approach. I will return to this topic in 6.2.2.

systematize the mentalistic terminology used in scientific psychology so as to ensure precision in experimental design and theory construction. I will return to this project in chapter 6, where I will examine various proposals and argue for what I think is the best approach: a synthesis of neuroimaging and conceptual analysis that will allow us to 'bootstrap' our way towards a more accurate cognitive ontology.

2.4 – Life After Universality

In section 2.2 I reviewed four distinct sources of evidence against the universality assumption: social cognition, anthropology, comparative linguistics, and experimental philosophy. Each of these suggests that there is sufficient evidence of variation in folk psychological discourse to at the very least render the universality assumption somewhat suspect. In 2.3 I considered various arguments that either claim that the use of such evidence is invalid, or try to account for it in ways that would not rule out the universality assumption. Whilst some of these arguments have merit, I think that ultimately we should still reject the universality assumption, at least in its strongest form. In the rest of this chapter I will consider the implication of accepting that there is genuine variation in folk psychological discourse. I will argue that folk psychological variation partially undermines appeals to intuition in philosophy of mind and cognitive science, and should lead us to exercise more caution when adopting common-sense 'mental' concepts such as belief and desire. The evidence of variation in explicit intuitions but not underlying mechanisms also provides incidental support for a two-systems approach to social cognition, which can help provide a motivated account of why we should keep these two phenomena (social cognition and folk psychology) distinct. Finally I will consider one reason why we might want to reject the 'no folk psychological miracles' argument even if folk psychological discourse did turn out to be universal. This will lead us neatly on to the next chapter, which will discuss how folk psychology can have a recursive impact on our minds and behaviour.

2.4.1 – Intellectual Humility

It seems fairly apparent that once we have accepted the evidence of variation in folk psychological intuitions, and rejected any argument for the superiority of our own intuitions, we should consequently exercise some humility when it comes to the use of intuitions as theoretical tools. The exact implications of this humility require some further unpacking, however.

Machery (forthcoming) considers three distinct ways in which we should be cautious about our use of philosophical intuitions: unreliability, dogmatism, and parochialism. If our intuitions are unreliable then we should be cautious of treating them as precise evidence, if they are dogmatic then we should be cautious as treating them as anything other than a restatement of the tradition that we have been educated in, and if they are parochial then we should be cautious about treating them as anything other than cultural idiosyncrasies. He takes the evidence from experimental philosophy to demonstrate that all three of these hold for philosophical intuitions, and that as a result "that we should suspend judgment in response to most philosophical cases" (*ibid*: 17).

With regard to folk psychological intuitions I have only really so far demonstrated that they are parochial, and that we should take care when generalising from our own intuitions to those of other cultures. What about dogmatism and unreliability? I think that the extreme focus on beliefs and desires in contemporary philosophy of mind might be a consequence of dogmatism. In undergraduate philosophy classes these tend to dominate the discussion, and coupled with Fodor's language of thought hypothesis they lead many to conclude that the only possible implementation of a computational theory of mind would be one that operated over beliefs and desires. In chapter 4 I will consider some cases where the focus on belief and desire has clouded philosophical and scientific discussion.

That folk psychological intuitions are unreliable might be demonstrated by considering the now well-established evidence of the impact that framing effects,

implicit biases[14] and the like can have on our intuitive judgements (see e.g. Brownstein 2016, Machery forthcoming). Given how widespread these appear to be, it would be unsurprising to find that they have an impact on folk psychological intuitions as well, which could lead us to make different judgements in different contexts (this would be an interesting avenue for future empirical research). If this were the case then we might want to adopt a general scepticism with regard to the reliability of folk psychological judgements.

2.4.2 – Two Systems Revisited

If folk psychology does turn out to provide an unreliable guide to the actual structure of cognition, then it might seem miraculous that we are able to achieve so much apparent success in our day-to-day social interactions (recall the 'no folk psychological miracles' argument discussed in 2.1.4). This is a common response to any attempt to revise or eliminate folk psychology, and it is somewhat plausible. However there is an easy way around it, which I have already indicated in my earlier distinction between social cognitive mechanisms and folk psychological discourse. Provided that we can give a principled story about the distinction between the two, then it is plausible that the former might enjoy predictive success and serve to guide social interaction without the latter conforming at all to how the mind actually works.

Luckily, a principled story of how to distinguish these two phenomena doesn't just exist, but is currently enjoying a fair amount of support in the empirical literature on social cognition. This is to adopt some version of a two systems theory of social cognition, where one system operates fast and implicitly in order to guide moment-to-moment social interaction, whilst the other operates slowly and explicitly, and might be responsible for what we actually say about other minds, even if that turns out to be false. I discussed various proposals along these lines in 1.3.2, but I briefly rehearse them here for convenience.

[14] The reliability and relevance of Implicit Attitude Tests has recently been questioned, but insofar as these are only one source of evidence for the existence of unconscious biases I take it that we should still exercise caution towards folk psychological intuitions.

The basic idea is that there is one system that rapidly processes the behaviour of others in terms of subdoxastic states such as "engagements" (Doherty 2011) or "registrations and encounterings" (Apperly & Butterfill 2009), alongside a second system that conducts a slower analysis in more familiar, folk psychological terms. These kinds of account were initially motivated by the discrepancy between developmental trajectories for explicit and implicit false belief tasks (discussed in the previous chapter). If the system that handles explicit reasoning develops slower than the implicit system, we might expect to see this kind of result. Typically the two systems are both taken to have the same target, but coupled with the mindshaping literature that I discuss in the next section and in chapter 3 I think they might allow us to account for a divergence between social cognitive mechanisms and folk psychological discourse. For most everyday situations the first, implicit system suffices, whilst the second system is only called upon for more complex or unusual situations, and will also be responsive to social and cultural factors that we might not normally think of as being about the mind itself. Cultural variation, then, would be a feature of system two but not (necessarily) of system one.

2.4.3 – A Self-Fulfilling Prophecy

Before moving on I want to discuss one final way in which we might undermine both the universality assumption and the no folk psychological miracles argument. This is the possibility, discussed in more detail in the next chapter, that it might be a mistake to think of folk psychology as being primarily in the business of *reading* minds – rather it might be better understood as a tool for *shaping* minds. Without going into too much detail here, the basic idea is that explicit folk psychological discourse might serve as a normative constraint on behaviour (both our own and that of other people), thus shaping the very minds that it purports to be describing. Even if folk psychological were universal, it still might not serve as an accurate guide to underlying cognitive mechanisms. Viewing folk psychology as a tool for shaping minds also allows us to neatly account for how folk psychological discourse could be predictively successful even if it does not track such mechanisms directly. By

shaping the cultural and social factors that constrain behaviour, folk psychology can create a 'niche' for itself where it becomes a self-fulfilling prophecy, successfully predicting behaviour not by being independently accurate, but simply by contributing to the very existence of that behaviour in the first place. A more in depth discussion of these possibilities is the topic of my next chapter.

2.5 – The Myth of a Universal Folk Psychology

This chapter has identified an assumption that folk psychology is universal, which is used to license the application of folk psychological concepts to the scientific study of cognition. This assumption was challenged on the basis that there is in fact evidence of significant variation in folk psychology, both across and within cultures. I considered various ways of responding to this evidence, and concluded that none of them are able to fully rescue the universality assumption, at least in its strongest form. So we are forced to face up to the fact that our own folk psychological assumptions are likely determined at least partially by cultural or linguistic factors, rather than by epistemic factors that would legitimate their role as a guide to the development of scientific concepts. By itself this does not rule out the use of folk psychological concepts in cognitive science, but it should at least begin to make us question our reliance on such concepts.

Chapter 3 – Folk Psychology as a Regulative Practice

In this chapter I will provide an account of the regulative role of folk psychology, drawing heavily on previous work by McGeer (2007), Zawidzki (2013), and Andrews (2015). This account will serve as the basis for an explanation of how folk psychology can continue to have predictive and explanatory success despite often being wrong about the underlying cognitive mechanisms. In the chapters that follow I will argue for the latter claim, that folk psychology is in fact often wrong about the underlying cognitive mechanisms, and consider what impact this might have on philosophy and cognitive science. For now I simply wish to defend myself against claims that folk psychology *must* be accurate given its predictive and explanatory success (see Fodor 1987 for the best example of this kind of claim). I will argue that the predictive and explanatory success of folk psychology can be explained, at least in part, by appealing to the ways in which folk psychology actively regulates and constrains our behaviour.

I will start in section 3.1 by introducing four distinct ways in which folk psychology might be thought of as 'successful' – these are related to the four components of folk psychological discourse that I identified in 1.3. I will then move on (in section 3.2) to consider the regulative role of folk psychology in more detail – here I will follow Zawidzki's (2013) account of 'mindshaping'. In section 3.3 I will relate this approach to more general work on cognitive niche construction (e.g. Clark 2006, 2008; Sterelny 2007, 2015), and argue that the regulative role of folk psychology qualifies as a kind of social-cognitive niche. Finally, in section 3.4 I will use this account of the regulative role of folk psychology in order to demonstrate how folk psychology can exhibit explanatory and predictive success without always being epistemically successful, where epistemic success is understood as successfully identifying underlying cognitive scientific mechanisms.

One thing that is worth noting before we proceed is that whilst previous discussions of the regulative role of folk psychology have tended to focus primarily on mental state attribution, I am assuming that folk psychology is a somewhat

broader practice (see chapter 1, especially section 1.3). So if folk psychology is able to operate in a regulative mode, it could do this in a number of ways, including mental state attributions, behavioural predictions, trait attributions, and the creation of narratives and schemas. This broader understanding of the regulative role of folk psychology is not entirely novel: Zawidzki, for example, emphasises the importance of behavioural predictions, and Andrews invokes something like a narrative when she describes how post-hoc explanations commit us to certain future behaviours. Nonetheless, the framework I describe here is novel in the sense that it brings all of this together for the first time, positing a single, unified role for folk psychology as a regulative practice.

3.1 – Four Kinds of Success

The success of folk psychology has traditionally been used to license an inference to the reality of the entities that it apparently posits. In the previous chapter I argued that evidence of cultural variation in folk psychological discourse undermines this inference, but my argument leaves the success of folk psychology unexplained. The primary aim of this chapter is to explain away the success of folk psychology without conceding that it accurately describes subpersonal mechanisms.

It is sometimes unclear exactly what people mean when they say that folk psychology is successful. In this section I will consider four distinct roles played by folk psychological discourse, and assess to what extent folk psychology can be said to 'succeed' in each of these roles. The roles in question correspond roughly to the four components of folk psychological discourse that I identified in 1.3; the predictive role corresponds to behaviour reading, the epistemic role corresponds to mental state attribution, the explanatory role corresponds to narrative competency, and the regulative role corresponds to normative constraints. The main focus of this chapter is on the regulative role, but I introduce the others here by way of contrast, and in order to illustrate the influence that the regulative role has on the success of each other role.

3.1.1 – The Predictive Role

The most basic way that folk psychology could be successful is by accurately predicting the behaviour of other people. In this role folk psychology is able to serve a useful pragmatic function, without necessarily being committed to any particular account of the hidden causes of the behaviour that it predicts. This is usually the kind of success that is appealed to when folk psychology is described as indispensable, although the assumption is typically that predictive success will rely on successful mental state attribution. Nonetheless, there are a number of different ways in which folk psychology could be predictively successful. One is indeed via the deployment of accurate mental state attribution, as is normally assumed, but this is certainly not the only way. Consider an everyday case: I am walking down the street and I see a smartly dressed person crouch down next to a bin and pick up a cigarette butt. I can predict with a fair degree of confidence that what they will do next is stand back up and drop it in to the bin. Whilst I could have made this prediction by attributing mental states and referring to an implicit theory of mind, it is also possible that I am just picking up on previously identified behavioural regularities, or perhaps explicitly situating this person's actions in a non-mentalistic narrative with which I am familiar. In any case successful behavioural predictions alone do not commit one to any further claims about the structure of cognition.

3.1.2 – The Epistemic Role

Another way in which folk psychology could be considered to be successful is if it gave an accurate description of how minds actually work. This is the kind of success that most people seem to have in mind when they cite the success of folk psychology as a reason for using it as the basis for scientific theories of cognition. If folk psychology was successful in this sense, then of course we would be justified in using it as the basis for our scientific theories, but by itself this argument is somewhat tautologous. Without some independent means of verifying the epistemic success of folk psychology, we can only justify using it as the basis for our scientific

theories once we have already established that the mind and brain actually do work in the ways described by folk psychology.

The direction of fit between folk psychology and how the mind actually works is important here. If folk psychology were correct because it accurately tracks how the mind works, then it would be a good guide to cognitive scientific discovery. However, if folk psychology were correct because it shapes the mind to fit the descriptions it provides, then it would only be a good guide to cognitive scientific discovery in those cases where it was able to shape the minds in question. The worry here is that the epistemic success of folk psychology might not generalise outside of those cases where there is a direct relationship between the folk psychologiser and the subject of their folk psychologising.

Historically this inferential move from the *predictive* success of folk psychology to the *epistemic* success of folk psychology was licensed by the (often implicit) 'no folk psychological miracles' argument that I discussed in the previous chapter (2.1.4). However, as I also established in that chapter, the argument is insecure given the apparent variation in folk psychological intuitions across cultures. It could be the case that some parts of folk psychology will turn out to accurately map on to actual cognitive mechanisms, but as we have no reliable way of identifying which these good parts are ahead of time, we should not rely on folk psychology for the identification of cognitive mechanisms.

3.1.3 – The Explanatory Role

The epistemic success that I described above is distinct from another important way in which folk psychology might be thought of as successful: its role as a provider of personal level explanations, rather than behavioural predictions or sub-personal models. To see how this role is distinct, recall the example of observing someone pick up a cigarette butt and predicting that they will put it in the bin. In addition to making this prediction I might want to explain why they did what they did. One kind of explanation would be to describe a sub-personal mechanism that caused their

behaviour, but as I suggested above (and in the previous chapter), we have no reason to think that folk psychology is especially good at giving explanations of this kind.

However there is another kind of explanation, perhaps equally important, that folk psychology is very successful at. I could explain the behaviour of this person by saying that they picked up the cigarette butt because it was the right thing to do, or because they hate litter, and so on. Explanations of this kind take place in what Sellars calls "the space of reasons" (1963: 169), and are sometimes called normative or rational explanations. They are distinct from sub-personal or causal explanations because they don't just describe why, in physical or mechanistic terms, the person behaved as they did, but they also provide a *reason* for this behaviour. Personal level reasons have a kind of explanatory traction that mechanistic reasons lack, especially when it comes to situating our behaviour in a wider socio-normative framework. Explanations of this kind are typically what people are after when they ask why someone did something, at least outside of the scientific context. Answering such why questions in terms of neurochemistry or physical causation somewhat misses the point, and is a domain in which folk psychology is almost indisputably superior to scientific psychology.

3.1.4 – The Regulative Role

There is a final, often neglected sense in which folk psychology can be successful: as a regulative practice, quite distinct from its success (or lack thereof) in each of the above three senses (although success in this regard can contribute to predictive and explanatory success, and vice versa). It is this regulative role that I will focus on in the present chapter.

The idea of folk psychology as a regulative practice has historical precursors in the work of Sellars, Davidson, and Dennett. Sellars' 'myth of Jones' implies, whether or not he himself intended it to, that the creation of internal mental states is a result of their external (linguistic) labelling, and Davidson picks up on this theme in his work when he argues that giving reasons for actions can in turn regulate future actions (1985). In his presentation of the intentional stance, Dennett (1987) indicates

that the predictive success of the stance is due partly to the fact that it also determines the limits of which actions are deemed rational. For both Davidson and Dennett, it is the presence of norms of rational behaviour that allows folk psychology to fulfil a regulative role. Below (and in the next section) I discuss how subsequent researchers have unpacked this idea and developed it in more detail.

McGeer (2007) was the first to discuss the regulative role of folk psychology in the context of contemporary social cognition.[15] She describes what she calls the "normative core" (*ibid*: 140-5) of folk psychology, which she thinks is implicitly present even in more standard accounts. This is the idea that there are some basic norms of rationality without which it would be impossible to apply folk psychological models across individuals. For example, if I predicted somebody's behaviour on the basis of attributions of certain beliefs and desires, I have to also assume that they connect these up in the right way, and are motivated to act so as to bring about whatever it is that they desire. Appeals to rationality of this kind are normative in the sense that there is a 'correct' way of acting, given a certain set of beliefs and desires, even if 'correct' in this case does not mean 'morally correct'. Without this basic normative core it is hard to imagine folk psychology ever developing beyond mere behaviourism, as in order to transcend predictions based on simple observations of inputs and outputs it is necessary to appeal to some broader set of rules that guides behaviour.

Zawidzki's recent formulation of the regulatory role of folk psychology as "mindshaping" (2008, 2013; cf. Mameli 2001) is explicitly inspired by Dennett's work. I will discuss it in more detail in the next section, but the basic idea is that we should re-orientate research in social cognition towards what he calls 'the mindshaping as linchpin hypothesis', which places the regulative role of folk psychology at the heart of social cognition (rather than the epistemic role, which he refers to as mindreading). He presents evidence for the primacy of mindshaping over

[15] A version of this proposal can be found in Mameli (2001), who coined the term "mindshaping" that was eventually taken up by Zawidzki, but McGeer was the first to address the idea systematically.

mindreading, based primarily on work in evolutionary and developmental psychology.

Matthews (2013), whilst focusing primarily on belief and other propositional attitudes, offers a brief account of the regulative dimension of folk psychology. He describes how "commonsense propositional attitude psychology may be predictively and explanatorily powerful […] through a process of enculturation [….] that ensures the predictive and explanatory efficacy of our culture's commonsense psychology" (*ibid*: 111). It is not entirely clear whether what he has in mind here is that commonsense psychology and culture might both have a shared cultural base, or that commonsense psychology might itself constitute that cultural base, but either way it seems that he is indicating the same kind of mechanism that I have in mind in this chapter.

Andrews (2015) has presented a more recent analysis of the regulative role of folk psychology, focusing on the connection between regulation and explanation. She argues that folk psychological explanations are also regulative, as they typically involve social interactions that invoke certain norms. For example, explaining why one was late to work in terms of forgetting to turn on an alarm involves presenting oneself as a certain kind of person, and perhaps committing oneself to not behaving in this way in the future (these are also the kinds of examples that McGeer and Davidson are concerned with). She describes the relationship between prediction, explanation, and regulation as a "tight spiral […] that modifies itself each time coordination breaks down" (*ibid*: 57), which captures the sense in which I think the regulative role of folk psychology contributes to its success in the other roles. Andrews also draws a connection between these kinds of looping effects and those described by Hacking, which I will return to in chapter 5.

Each of these approaches contributes to what I am calling the regulative role of folk psychology, the role in which it is able to guide behaviour by shaping the socio-normative framework within which we live our lives. At its simplest this framework merely requires that we maintain internal coherency in our actions, so that once we have committed ourselves to believing something, we ought to act in

such a way that is consistent with believing that thing. At the more complex end we are forced to give full explanations for our behaviour, on pain of seeming irrational, or in some cases even immoral. And once an explanation has been given it creates a new narrative that we must maintain consistency with. In this way folk psychology is able to create the very systems that it describes, at once predicting and making true that very same prediction. In the rest of this chapter I will discuss in more detail the mechanisms that enable this regulative role, and consider the implications that it has for our understanding of folk psychological discourse.

3.2 – Varieties of Mindshaping

In this section I will examine the regulative role of folk psychology in more detail, following the template set by Zawidzki's taxonomy of the "varieties of mindshaping", which include "imitation, pedagogy, norm cognition and enforcement, and language based regulative frameworks" (2013: 29). I will describe each in more detail below, and explain how they relate to my understanding of folk psychological discourse.

Prior to describing each variety of mindshaping, Zawidzki formulates a general definition of the phenomenon. Mindshaping is defined in terms of three components and a mapping relation: a model, a target, a mechanism, and the sense in which the target is intended to match the model. The mechanism's function is to shape the target to match the model (in the relevant respects). The target is a mind of some kind, be it the shaper's own or someone else's. The model may be another mind, some representation in the mind of the shaper, or even something more abstract like an idealized set of behaviours or a fictional character. The sense in which the target is intended to match the model might be more or less precise, depending on the mechanism in question. Note that these terms are all intended to be used in a very general sense, and need not necessarily imply any intentionality or agency on behalf of either the target or the model. (Adapted from Zawidzki 2013: 31-2.)

To give a paradigmatic example, mindshaping might occur when an agent

(X) issues a command to a conspecific (C). That command serves as a mechanism for shaping C's mind, taking advantage of cognitive processes that X might not even be aware of. A verbal command is issued by X, the words are processed by C, and C consequently feels a social pressure to conform to X's command. In this case C is the target, the model is whatever behaviour X wants C to perform, and the details of that model being the sense in which C's behaviour should match X's command. In this example both parties are explicitly aware of what is going, but other cases of mindshaping can be much more subtle, and can even be self-directed, i.e. the target, model, and mechanism can all be contained within the same agent. I will now consider several distinct mechanisms by which mindshaping might be achieved.

3.2.1 – Imitation

Both human and non-human primates, as well as some species of birds such as crows, engage in relatively sophisticated imitation that qualifies as a kind of mindshaping. For example, human infants are known to imitate the facial expressions and verbalisations of their caregivers, and chimpanzees are able to quickly imitate novel problem solving behaviours. In both of these cases the target system is the infant or chimpanzee themselves, the mechanism is whatever cognitive system enables imitation, the matching relation is whatever degree of accuracy they are able to copy the behaviour, and the model is either the caregiver or the problem-solver, respectively.

One proposed mechanism for imitation is the mirror-neuron system, which consists of motor neurons that activate not only when performing an action, but also when observing a similar action. Whilst the evidence for the existence of such a system in humans is somewhat controversial (see 1.2.2), it could potentially serve as a basic mechanism for imitation by connecting observed behaviours with potential actions. Zawidzki summarises imitation thus:

> In imitation, the targets are the imitator's mind or behavioral dispositions; the model is another concrete, nonfictional individual; the mechanism is some pattern of activity in the imitator's nervous system,

> possibly involving mirror neurons; and the respects in which the target is shaped to match the model correspond to properties of model behavior to which imitation mechanisms (e.g., mirror neurons) are sensitive. (Zawidzki 2013: 42)

Zawidzki distinguishes between non-human imitation, which appears to always be extrinsically motivated (i.e., the imitation is always a means to an end), and human imitation, which can be intrinsically motivated (2013: 35). This leads human imitators to sometimes copy irrelevant behaviour, resulting in a more liberal mapping relation than in non-human imitators. In one study (Horner & Whiten 2005), the performance of young chimpanzees and human infants was compared on learning to open a simple puzzle box, which was either transparent (allowing the participant to directly observe the mechanism) or opaque. In each case the participant observed an experimenter solving the puzzle, but also performing an irrelevant additional step. In the opaque condition (where it was not obvious that this step was irrelevant), both humans and chimpanzees performed both steps. However, in the transparent condition the chimpanzees performed only the relevant step, whilst the human infants continued to perform both steps, perhaps indicating an intrinsic motivation for imitation. Human children also routinely over-imitate their adult caregivers, copying not only functionally relevant features of the behaviour in question, but also irrelevant features such as the speed or style in which an action is carried out.

The human capacity for intrinsically motivated imitation might go some way towards explaining the predictive success of folk psychological discourse. If we frequently (and perhaps unwittingly) imitate even non-functional features of each other's behaviour, we will tend towards behavioural conformity, which constrains the range of possible predictions. Zawidzki also emphasises that both human and non-human imitation does not seem to require sophisticated mindreading, as it primarily involves the imitation of *behaviours* rather than mental states (i.e. the imitator need not have any understanding of *why* their target is doing what they are

doing). Thus behavioural predictions that rely on imitation-based conformity could succeed regardless of the epistemic status of folk psychology.

For example, I can predict that once I yawn, many other people in the same room as me will also begin yawning, without possessing any knowledge or understanding (whether implicit or explicit) of the cognitive mechanism responsible for 'contagious yawning'. This is because I am sensitive to a behavioural regularity that is the result of low-level imitation. The point here is that my sensitivity to the regularity itself is sufficient for me to successfully predict this behaviour, without requiring that I understand how the regularity comes about. So low-level imitation can enable basic behavioural predictions without invoking high-level mindreading of any kind.

3.2.2 – Pedagogy

Another interesting kind of mindshaping is pedagogy, i.e. any explicit transmission of skills or knowledge from one agent to another. Unlike imitation, the mechanisms involved in pedagogy span both the learner and the instructor, who in some cases may also be the target.

For many readers, the most familiar example of pedagogy will be classroom instruction of some kind, but as Zawidzki notes, this "is probably a relatively recent and atypical form of pedagogy" (2013: 43). For much of human history it is more likely that pedagogy was primarily a small-scale affair, with novice learners observing the behaviour of experts practising their craft (cf. Sterelny 2007, 2012). This suggests a more gradual progression from imitation to pedagogy, with many forms of pedagogy essentially consisting of structured imitation. The connection between imitation and pedagogy is also strengthened by the observation that human infants also tend to overgeneralise in pedagogical situations, such as when they apply specific grammatical rules to cases where they should not apply.

For my purposes one especially interesting kind of pedagogy would be that which involves the explicit teaching of folk psychological behaviours and concepts. For example, we could imagine a society in which children are taught the basics of

belief-desire psychology at school,[16] and so interpret each other's behaviours according to this framework, regardless of how accurate it actually is. Much like in Sellars' myth of Jones, the explicit adoption of this interpretive framework could lead to people's behaviours eventually conforming to it, even it were not originally true (or even if it never came to reflect the underlying mechanisms driving that behaviour). More plausibly, consider the kinds of narrative explanation that I described in 1.3.4. Infants typically learn such narratives from their caregivers, and are likely to conform to the roles and structures that they describe. So narratives serve a dual function, both regulating our own behaviour whilst also allowing us to explain the behaviour of others (see 3.2.4).

3.2.3 – Norm Cognition and Enforcement

The most important kinds of mindshaping, at least so far as this chapter is concerned, are those that explicitly facilitate behavioural conformity, and thus enable folk psychology to achieve greater predictive and explanatory success than imitation and pedagogy would alone. Zawidzki distinguishes two classes of mindshaping mechanisms of this kind: one unconscious and automatic, and the other more explicit. He also discusses the regulative role of language, which I will cover in the next section. Whilst these mechanisms are related to imitation and pedagogy, they differ in that their primary function is to regulate social behaviour rather than achieving some other non-social benefit (such as learning a non-social skill like fishing).

An example of an unconscious and automatic mindshaping mechanism is the so-called "chameleon effect", reported in classic studies by Chartrand & Bargh (1999; discussed by Zawidzki 2013: 50-3). In these studies subjects collaborated with confederates on a simple task, whilst the confederates performed several non-functional body movements, such as crossing their arms or smiling. The subjects frequently mimicked these movements, despite being seemingly unaware of doing

[16] In a dystopian future where Fodor rules supreme and the Churchlands have been driven underground.

so. Whilst perhaps not strictly folk psychological, the cognitive mechanisms behind such effects (and related priming effects) might serve as the basis for the regulation of behaviour. Chartrand & Bargh suggest that the chameleon effect acts "as a kind of natural 'social glue' that produces empathic understanding and even greater liking between people" (1999: 897), which could enable more complex social cognitive phenomena such as joint attention and co-operation.

More explicit or intentional forms of social norm enforcement include a wide range of behaviours, but Zawidzki focuses on those that involve some form of punishment. People are generally willing to exert a fair amount of effort (at no immediate benefit to themselves) to punish others who break certain norms, such as fair resource distribution (see e.g. Henrich 2009; Henrich *et al* 2005, 2006, 2010; discussed by Zawidzki 2013: 53-4). More informally, we may think of social norms such as good manners, as well as various moral norms, as being enforced by costly punishment of some form or another. In terms of folk psychological regulation, enforcement of the norms of rationality described by Davidson and Dennett would tend to result in more homogenous, and thus more easily predicted, behaviour.

To give a simple example, consider what happens when a fire alarm goes off in a building. Most people will leave the building immediately, unless they are either unaware of the meaning of the alarm, unable to hear it, or else aware that the alarm is tested every Thursday morning, and so should be ignored. If we saw someone sitting in their office whilst everyone else was leaving, we would be concerned about their wellbeing, and if they persisted in not leaving even after we had knocked on their door and caught their attention, we might begin to think a number of different things about them: perhaps they have a death wish, or perhaps they are stubborn and stupid, or perhaps they are simply completely irrational. In each case we might have reason to shun or censure them on future occasions, as their failure to conform to established behavioural norms potentially puts other people in danger. So over time we should expect everyone to conform to the 'fire alarm norm', making behaviour relatively easy to predict whenever a fire alarm goes off.

3.2.4 – Language Based Regulative Frameworks

Finally, mindshaping can be achieved via explicit linguistic narratives, which can serve on one hand to regulate the range of possible behaviours, and on the other hand to make sense of the behaviour of others within a familiar framework. Zawidzki notes that the focus of previous work on the role of narratives in social cognition has been primarily in the latter direction (see e.g. Hutto 2008), but argues that the former role is also very important (Zawidzki 2013: 57-61). As Zawidzki puts it, narratives not only help us make sense of behaviour, but they also "help constitute the minds that such knowledge enables us to track." (*ibid*: 57). They do this through what Zawidzki describes as "self-constituting narratives", which are narrative roles that become internalized by an individual and govern their behaviour. For example, a person might internalise a narrative about the importance of individual striving for success, leading to them conceiving of themselves as a strong, individualist loner, and thus behaving as such. Someone else who was familiar with this kind of narrative might be able to pick up on behaviour cues that, with the help of the narrative, would allow them to predict this individual's behaviour. (Andrews 2015 describes this process in more detail.)

3.3 – Folk Psychology as a Cognitive Niche

In this section I will argue that a useful way of characterising the regulative role of folk psychology is as a form of cognitive niche construction (cf. Clark 2006, 2008; Sterelny 2007, 2015). The varieties of mindshaping that I described in the previous section each contribute to the creation of a 'folk psychological niche' that regulates human behaviour and functions, amongst other things, in order to make it easier to predict and explain.[17]

[17] In *The Cultural Construction of Belief* (unpublished draft), Matthews will make a similar claim, "that our commonsense psychology is a culture-specific niche construction that serves a primarily normative/regulative role" (see http://www.robertjmatthews.org/work-in-progress.html). I am not yet sure what the details of his account will be, and in what ways (if any) it will differ from mine.

3.3.1 – Cognitive Niche Construction

The idea of a cognitive niche is based on earlier work on niche construction in ecology (see e.g. Laland, Odling-Smee, & Feldman 1999, 2001; Odling-Smee, Laland, & Feldman 2003). An ecological niche is simply the environment within which an organism lives, usually with the implication that it is to some extent adapted to that environment. Niche construction is the phenomenon of an organism modifying its environment in order to make that environment more suitable for it or its conspecifics. Perhaps the most famous example of this is the construction of dams by beavers, which creates a (relatively) safe environment within which they can live and raise their young. Other examples include the regulation of soil chemistry by earthworms, tool use in primates and some corvids, and, of course, the multitudinous ways in which humans structure their environments.

Human niche construction, perhaps uniquely, involves not only physical artefacts and environmental changes, but also abstract cultural 'tools', such as language, (arguably) morality, and, as I will argue, folk psychological discourse (which is itself part constituted by linguistic practices). Language is especially important here, and is perhaps essential for the kinds of cognitive niche construction that seem to be uniquely human. As Clark puts it, "language (and material symbols more generally) [provide] a new kind of thought-enabling cognitive niche" (2006: 370), which opens up novel problem solving strategies and accelerates human cultural evolution. Clark (and others) have focused primarily on the cognitive niche constructed by the internalisation of natural language, but in the next section I will suggest that another way language contributes to niche construction is via the mindshaping mechanisms that I described previously. Zawidzki (2013: 128) also makes this connection, and indicates that human cognitive niche construction might differ from other kinds of niche construction in that humans can purposefully choose how they shape their environments.

Consider an example that Clark gives, of how advanced number cognition might be facilitated by the creation of linguistic labels ('one', 'fifty-four', etc.) that

allow us to categorise and simplify the numerical domain. He cites research where a chimpanzee, Sheba, was taught to use numerals that enabled her to succeed in a counter-intuitive task where picking the larger pile of food would actually result in her being given less food. Without the numerical labels she was unable to pass this task, and would continually pick the larger pile, but once a numerical label was attached to each pile she was able to make sense of what was going on, and pick the smaller pile. Thus it seems that the numerical symbols helped simplify the task domain, creating a cognitive niche that facilitated more adaptive behaviour. (Drawn from Clark 2006: 371, originally study by Boysen *et al* 1996).

Clark hypothesises that human language might operate in much the same way as numerical symbols did for the chimpanzee, simplifying an otherwise overwhelmingly complex environment by attaching labels to especially relevant features. Something similar, I want to suggest, could be the case for folk psychology.

3.3.2 – The Folk Psychological Niche

In the previous section I described how some animals, including humans, are able to construct their own environmental niches that contribute positively to their welfare. One particularly interesting kind of niche construction involves the creation of *cognitive* niches, i.e. environments that contribute positively to cognition. Folk psychological discourse, I will now claim, constitutes a cognitive niche that contributes to its own predictive and explanatory success. It does this by regulating the behaviour of conspecifics (and oneself), and by creating a socio-normative framework within which rationalistic explanations can be given.

Each of the mindshaping mechanisms that I described earlier in this chapter contribute to the construction of the folk psychological niche. Imitation and pedagogy both tend to normalise behaviour by transferring certain behavioural patterns between individuals. For example, if everyone is taught how to make a cup of tea in a particular fashion, then tea-making behaviour will consequently become easier to predict. Similarly, both linguistic and non-linguistic norm enforcement also regulate behaviour, by 'punishing' atypical behaviour, either literally or via social

mechanisms such as exclusion or ostracisation. More constructively, the transmission and propagation of narratives creates a framework within which to situate and explain behaviour (cf. Andrews 2015).

Let us focus for a moment on the particular example of pedagogy. Both formal classroom instruction and the master-apprentice scenario described by Sterelny (2012) involve the regulation of behaviour by reinforcing whatever behaviour is being taught. Whilst the reinforcement mechanisms might differ across cases, and be more or less successful, the end result that is aimed at is producing students who can replicate the behaviour being taught – which might simply be repeating facts, or might be a more complex skill such as woodwork or fishing. In each case, provided that we are situated in the same pedagogical framework, it will become easier to predict the student's behaviour when they are engaged in the relevant practice. So provided that I was taught in the same school of fishing as you, it should be fairly easy for me to predict the order that you prepare your tackle, and so on. Whilst there will inevitably be exceptions to these behavioural regularities, they will not be the norm, and over time will be corrected by social pressures (or in the case of exceptions that prove to be more successful than the norm, they may end up spreading and becoming the new norm). Teaching someone to fish not only provides them with a useful skill, but also contributes (in a small way) to the creation of a stable folk psychological niche, in the sense that we will now be better equipped to predict what the student will do when they pick up a fishing rod (and so on).

Of course, pedagogy (and imitation) is only one part of the picture. Perhaps the most interesting mindshaping tool for the purposes of folk psychological niche construction is the regulative role played by folk psychological attributions themselves. The very act of ascribing a belief or desire to another (or to oneself) creates a social pressure to conform to this ascription, as otherwise one would undermine either one's own or someone else's rationality. In this way folk psychological discourse can become a self-fulfilling prophecy, as the seemingly epistemic claims it makes sometimes end up creating the very behaviours that they predicted. There is a further question here about whether or not this means they

actually are accurate, but at the very least the creation of a cognitive niche can help explain the predictive and explanatory success of folk psychological discourse.

In a sense this is just an elaboration on "the game of giving and asking for reasons", initially described by Sellars (1963) and later elaborated on by Brandom (1994). By attributing mental states to one another, we create a 'space of reasons' that constitutes a kind of cognitive niche within which we can expect people to behave in certain systematic and relatively easy to predict ways. For example, if I tell you that the meeting tomorrow is cancelled, I can expect you not to turn up for it, based on the norms of rationality that are partly constituted by this cognitive niche. Our explicit commitment to these norms could facilitate systematic predictions, even if the underlying cognitive architecture was unchanged.

3.4 – Failing With Style

In this final section I will describe how folk psychology can often give a strictly false account of how the mind works (at least in terms of cognitive scientific mechanisms), whilst nonetheless remaining a successful predictive and explanatory practice. It does this by cultivating a cognitive niche, as described in the previous section, which regulates behaviours and thus mitigates the computational intractability of behavioural predictions. This creates a socio-normative 'space of reasons' within which explanations of behaviour can be given with relative ease. Finally, the niche also contributes to whatever epistemic success folk psychology does have, via shaping cognitive mechanisms themselves and thus becoming a sort of self-fulfilling prophecy.

3.4.1 – Epistemic Failure (and Occasional Success)

In the second half of this thesis I will argue that folk psychology often fails at the epistemic role, at least so far as the fine grained structure of sub-personal mechanisms are concerned. I will briefly rehearse these arguments here, but motivating this claim is not the main aim of this chapter. By claiming that folk psychology fails in the epistemic role, I do not mean to say that it never gets

anything right, or even that it is a 'false' theory in the sense meant by the Churchlands, but rather that it does not provide a reliable guide to the sub-personal structure of cognition. Even if it is occasionally successful at describing the sub-personal mechanisms responsible for cognition, we have no good reason to think that this is typically the case, and as such it should not be relied upon as a guide to scientific discovery. As I will argue in the remainder of this section, failure of this kind is perfectly compatible with folk psychology succeeding at (personal level) prediction and explanation.

Previously, in chapter 2, I have argued that evidence of cultural variation in folk psychological intuitions gives us an initial reason to be sceptical of the epistemic value of folk psychology. This is because an implicit assumption that folk psychology is universal has historically been used to license the use of folk psychological intuitions as a guide to actual cognitive structures. Once this assumption has been proved false, this usage is no longer so obviously valid.

In chapter 4 I will present several case studies from philosophy and cognitive science, each of which demonstrates a way in which folk psychological concepts fail to capture the complexity of cognitive scientific explanation. Typically what is needed in these cases is a disambiguation between different senses of a single folk concept, or sometimes a recognition that the distinctions made by folk psychology do not straightforwardly map on to the functional structure of actual cognitive systems.

Finally, in chapter 5 I will challenge the status of folk psychological kinds as genuine natural kinds, and argue that folk psychological discourse fails to provide genuinely projectable predicates. This is partially motivated by the evidence from chapters 2 and 4 – folk psychological kinds are culturally variable, and also cross-cut relevant functional and structural distinctions in cognitive science. However, folk psychological kinds may constitute culturally mediated 'human kinds', via the mindshaping mechanisms explored in this chapter. I will return to this topic in 3.4.4, and again in 5.2.4.

It is important to note that regulative success could also contribute to the occasions when folk psychology is able to correctly identify sub-personal

mechanisms. The varieties of mindshaping described in 3.2 not only shape behaviour, but sometimes also shape the cognitive mechanisms responsible for that behaviour. Presumably all behaviour involves a cognitive mechanism at some stage, but what I am especially interested in here are those cognitive mechanisms that come to closely resemble the folk psychological attributions that purport to identify them. This requires us to distinguish between two ways in which different kinds of mindshaping could operate. One way would be to make only surface level changes to the content of people's cognition, for example by propagating certain ethical or rational norms. Another way would be to actually change the sub-personal mechanisms involved in cognition, which we might think is the outcome of certain kinds of pedagogy. In the latter case, but not the former, mindshaping could result in cognitive mechanisms that conform to the descriptions given by folk psychological discourse. When this occurs folk psychology will be epistemically successful, but in virtue of it shaping the mechanisms it identifies, rather than being especially good at identifying those mechanisms independently. I will return to this topic in chapter 5, where I discuss the self-constituting nature of folk psychological kinds.

What I want to suggest is that folk psychological discourse may be potentially flawed as an epistemic tool for learning about the sub-personal structure of cognition, even if from time to time it does get something right. However, this does not mean that it should be eliminated (or even revised outside of scientific contexts). Rather, what I take it to mean is simply that conceiving of folk psychology as serving this epistemic role was a mistake in the first place, and that we would do better to focus on the predictive, explanatory, and regulative roles of folk psychology.

3.4.2 – Regulative Success

As I described in the previous sections, folk psychological discourse exerts a remarkable range of regulative and normative pressures on our behaviour and cognition, via a class of mechanisms described by Zawidzki as 'mindshaping' (in contrast to the epistemic role implied by 'mindreading'). One simple example of this

is the kind of pressure that one might feel to conform to rational implications of previously expressed beliefs. So if you are (publicly) committed to the belief that Edinburgh is north of Paris, then you might feel pressure to head south rather than north if you tell someone that you are travelling from Edinburgh to Paris – unless you are able to provide some additional explanation for your actions, such as that you are catching a flight from Aberdeen. This pressure need not necessarily be public, either – even becoming aware of one's own folk psychological commitments might be sufficient to rationally constrain one's own actions.

This is only one small way in which folk psychology can exert a regulative influence (see section 3.2 for more detail). Zawidzki gives many other examples, including imitation, pedagogy, and more explicit norm enforcement (such as ethical discourse). If we accept that these all count as cases of folk psychological regulation, then it certainly seems that folk psychology is successful in this role. However, it is important to note that what counts as 'success' here is going to be somewhat contextual – formal pedagogy has institutionalised measures of success in the form of grades, assessments, etc., whilst what counts as the successful enforcement of norms of rationality might be somewhat more open to interpretation. In the above example I noted that one possible response to contravening a rational norm is to offer an alternative explanation of one's actions. Andrews (2015) makes this point as well, but goes on to argue that giving an alternative explanation nonetheless entails further commitments – once you have established that you are heading north to go the airport, your future behaviour can be expected to involve attempting to get on to a plane, and so on. When this game of giving rational explanations breaks down (either with regard to the behaviour or the explanation) we tend to resort to attributions of mental illness or other cognitive disturbances. These serve almost as a 'trump card' explanation that can account for any behaviour at all, although the extent to which they are genuinely explanatory (rather than just excusing one from giving explanations) could be challenged.

In any case, folk psychology certainly seems to succeed in the regulative role in many everyday situations. It is this success, I want to suggest, that can account not

only for the explanatory and predictive success of folk psychology, but also many apparent cases of epistemic success. When folk psychology is epistemically successful, I argue, this is typically not due to any particular competency it has at identifying sub-personal mechanisms, but rather due to the influence that the regulative role sometimes plays in the *formation* of sub-personal mechanisms. Essentially, in certain cases folk psychology can act as a self-fulfilling prophecy, making it no miracle that (in these cases) it correctly identifies the mechanisms involved.

3.4.3 – Explanatory and Predictive Success

Regulative success can also help explain the predictive and explanatory success of folk psychology. The creation of a homogeneous behavioural environment contributes to predictive success, and the enforcement of rational norms supports a certain kind of explanatory practice. This helps us to account for how folk psychology can exhibit predictive and explanatory success without necessarily exhibiting epistemic success, thus further undermining the 'no folk psychological miracles' argument outlined in the previous chapter. I will now discuss the contribution made by regulative success to predictive and explanatory success in more detail.

Zawidzki (2013: chapter 3) describes how the complexity of human behaviour provides a problem for the view that folk psychology primarily consists of mindreading (what I call the epistemic role):

> Because any observable behavior is compatible with any finite set of propositional attitudes, accurate propositional attitude attribution that is timely enough to make a difference to behavioral prediction in dynamic, quotidian contexts appears to be computationally intractable. (Zawidzki 2013: 65)

The same problem applies to behavioural predictions. Any observable behaviour is compatible with many future possible actions, and even if it was theoretical possible to predict which action would come next, it might be practically impossible to

actually make that prediction quickly enough for it to be useful. Furthermore, both future behaviours and mental states seem to be routinely underdetermined by current behaviour observations, as described by Dennett (1987) – although Dennett goes on to argue that this indeterminacy may simply be a feature of cognition, not just a problem with folk psychology. Based purely on observation, both epistemic and predictive successes seem impossible.

If Zawidzki is right, then human behaviour is by default simply too heterogeneous for predictive success. However, we are clearly able to predict each other's behaviours with a relatively high rate of success, at least under normal circumstances. The trick, according to Zawidzki, is that folk psychology acts in the regulative role so as to constrain human behaviour and make it relatively homogenous and easy to predict. It does this via the various mindshaping mechanisms described in the previous section, such as basic imitation serving to homogenize behaviours, and pedagogy and norm enforcement teaching people which kind of behaviours are appropriate in what kind of situations. So long as people are working under the same set of folk psychological norms, otherwise intractable predictions should become relatively easy to make.

What about explanatory success? Andrews (2015) describes how folk psychological regulation can contribute to this by creating looping effects where the giving of folk psychological explanations serves to create future constraints on behaviour, and where those constraints consequently limit the kinds of explanations that can be given. She describes a 'folk psychological spiral' that also includes predictive success, which becomes interwoven with explanation and regulation. This picture of folk psychology allows us to separate predictive, explanatory, and regulative success from epistemic failure, and gives a principled foundation on which to build a non-epistemic account of folk psychology. Andrews give a helpful example of how this kind of explanatory constraint could work in practice:

> The folk psychology spiral works dynamically in our more complicated social interactions. Suppose Ernie and Bert are trying to decide where to

> go on holiday, and Ernie is lobbying for a trip to Bali. Bert is not convinced that it is a good idea, but Ernie predicts that Bert will love Bali, because he enjoys the arts and the outdoors and monkeys and tropical fruit. Bert is persuaded, but says to Ernie, "You'd better be right about this!" They arrive in Bali and Ernie's prediction is born out—Bert has a great time. That reinforces Ernie's belief in Bert's preferences, and also creates added pressure for Bert to live up to them—Ernie says, "See, I told you you'd love Bali!"
>
> But if Ernie's prediction fails, and Bert doesn't enjoy Bali, Ernie (wanting to preserve the relationship) will seek to figure out why. Maybe Bert hates the tropical heat more than he likes the outdoors, monkeys, fruit, and arts. That explanation leads Ernie to form additional expectations about Bert's preferences, and so Ernie might book an air-conditioned hotel. Ernie would expect Bert to be happier with the air-conditioning, and this expectation creates a pressure for Bert to appreciate the air-conditioned room, to express thanks to Ernie for considering his needs, and so forth. (Andrews 2015: 57-8)

So in giving an explanation of why Bert didn't enjoy the holiday, Ernie creates future pressures on Bert to behave in certain ways, further constraining the space of possible behaviours. This in turn will lead to further cases of prediction and explanation, forming what Andrews describes as a 'folk psychological spiral' of prediction, explanation, and regulation, all of which contributes to future predictive success.

3.5 – The Regulative Role of Folk Psychology

In this chapter I have considered four distinct roles that folk psychology might succeed or fail at. In the predictive role folk psychology simply predicts future behaviour based on past behavioural observations. In the epistemic role folk psychology aims to describe the actual mechanisms and processes that enable cognition. In the explanatory role folk psychology provides (often post-hoc) explanations for peoples' behaviours. Finally, in the regulative role that is the focus of this chapter it is able to exert a regulative influence on behaviour by establishing social, rational, and ethical norms. I then argued that success in the regulative role enables success in the predictive role, explanatory role, and to some extent the epistemic role. This explains how folk psychology can function as a successful social

practice despite often failing to identify genuine sub-personal mechanisms, thus further undermining the 'no folk psychological miracles' argument presented in the previous chapter.

In the second half of this thesis I will present further evidence that folk psychological discourse is not suitable for technical work in philosophy and cognitive science (chapter 4), consider the status of folk psychological kinds as natural kinds (chapter 5), and discuss how we might go about replacing folk psychological concepts in cognitive science (chapter 6). First, though, I will briefly pause to recap the positive account of folk psychology presented in the first half of this thesis, describe how it all fits together, and consider both how it differs from more traditional accounts, and how it relates to other alternative accounts.

Interlude: The Positive Account of Folk Psychology

So far in this thesis I have presented an overview of different uses of the term folk psychology (chapter 1), described how folk psychology appears to vary across cultures and what the implications of this variance might be (chapter 2), and considered how folk psychology might sometimes be able to operate in a 'regulative role', exhibiting an active influence on our future behaviour (chapter 3). In this section I will try to present more clearly the positive account of folk psychology that I see as emerging from these reflections. This positive account has important antecedents, perhaps most obviously in Dennett's work on the 'intentional stance' (see his 1987), but also in other works such as Schwitzgebel's dispositional account of belief (2002), and Andrews' call for a "pluralistic folk psychology" (2008). At the end of the section I will discuss how my account builds upon and advances these previous proposals.

At the heart of my account is the proposal that we should distinguish between implicit social cognitive mechanisms on the one hand, and explicit folk psychologising on the other – and the thought that once we have done so, many further puzzling features of the general phenomenon referred to as 'folk psychology' can be explained. The most important of these for my purposes is that we can make sense of how the explicit content of folk psychology could fail to match up with our scientific understanding of cognition, whilst nonetheless still managing to track important and interesting features of human behaviour. The way to do this, I think, is to see explicit folk psychological attributions as descriptions of traits or dispositions of whole people, rather than as attempts to pick out discrete mental states, processes, or mechanisms. So when I say that someone believes something, I am saying that they will act in ways that rationally accord with believing that thing, rather than trying to say anything at all about the cognitive mechanisms that cause them to act in that way. Attributions of belief, for example, need not be committed to the actual existence of any belief-like state, but can be 'true' merely in virtue of someone behaving *as though* they believed that thing.

At the same time, it may not be the case that implicit social cognitive mechanisms actually attribute beliefs at all. They might track far simpler behavioural regularities that suffice for making many of the social predictions that we require on a day-to-day basis, whilst falling back on explicit folk psychology for explanations or predictions of more complex or unusual behaviour. Making this distinction means that we can assess the success or failure of explicit folk psychological attributions separately from more basic social cognitive competency, and can help to make sense of the evidence of cross-cultural variation that I discussed in chapter 2. Furthermore, the regulative mechanisms that I discussed in chapter 3 (which can operate both implicitly and explicitly) help to explain how folk psychology could successfully pick out traits or dispositions without successfully identifying discrete cognitive states. The idea is that whilst behavioural prediction can only get us so far, it will be able to get us further if we are able to shape the systems that we are predicting, i.e. human cognitive systems, so as to make their behaviour more regular. So by shaping our social cognitive environment we are able to create a 'niche' of sorts, within which behavioural and otherwise non-mentalistic strategies can successfully predict and explain a lot of human behaviour.

In the second half of this thesis I will go on to assess the status of folk psychological concepts in scientific psychology and cognitive science, and the picture of folk psychology outlined above will allow me to do so without worrying about whether the failure of folk psychological concepts in this regard would have to lead to their eventual elimination. Even if it turns out that there is nothing like a discrete belief state anywhere in the brain, folk psychology can quite happily carry on referring to beliefs as character traits or dispositions, and so on for any other folk psychological concept (see Botterill & Carruthers 1999 for a similar response to eliminativist arguments).

The idea that folk psychology should be interpreted as describing coarse-grained behavioural states rather than fine-grained mental states has an obvious antecedent in Dennett's intentional stance. Dennett characterised three distinct 'stances' we could adopt when explaining or predicting the behaviour of a system.

The physical stance looks at the actual physical structure of a system, the design stance looks at the intentions of a designer (or imagined designer) of a system, and the intentional stance treats that system as though it has its own intentions, and uses these to predict its behaviour. Importantly for Dennett, we can adopt the intentional stance towards systems whose actual mental states we know nothing about, or even towards a system that we would not normally treat as mental (such as a thermometer). Dennett can be interpreted as either saying that we should be instrumentalists about folk psychological attributions, or that such attributions actually apply to the behaviour of whole systems, rather than parts of those systems. In a sense what I am doing in this thesis is adopting that latter interpretation, and trying to see how far we can take it (and what implications, if any, it has for experimental cognitive science).

The idea that folk psychological concepts should be characterised as dispositional can be found in Schwitzgebel (2002), although he focuses only on belief. Schwitzgebel argues that belief should be characterised in terms of behavioural, cognitive, *and* phenomenal dispositions, i.e. dispositions to behave in certain ways, dispositions to think certain things, and dispositions to have certain experiences. He supports this argument by describing a number of cases where beliefs do not seem to have the discrete, determinate nature required of mental states as classically understood. Whilst I have focused primarily on behavioural dispositions, I am sympathetic to Schwitzgebel's approach, and seek to extend this kind of analysis to folk psychology more generally.

Finally, Andrews (2008) outlines a pluralistic attitude towards folk psychology, where trait attribution can play some role in prediction and explanation, alongside the more traditional mental state attribution. Trait attribution is the attribution of personality traits, such as being greedy or brave, to whole people, as opposed to the attribution of discrete mental states such as belief and desire. I agree with Andrews that trait attribution is an important and often ignored component of folk psychological discourse, and coupled with the dispositional account of mental state attribution described above I think her pluralistic approach has a lot in common

with my own account. Andrews also spends some time identifying and criticising the emphasis that mainstream philosophy of mind has historically placed on propositional attitude attribution. She focuses primarily on the implications her approach has for social cognition, but I have tried to apply my account to folk psychology more generally. In the following chapters I expand this analysis to the use of folk psychological concepts in cognitive scientific discourse.

Chapter 4 – Folk Concepts in Cognitive Scientific Discourse

In this chapter I will present several cases where the application of folk psychological concepts to issues in philosophy and cognitive science appears to lead to practical difficulties and/or theoretical confusions. In each case I will suggest that a strategy of conceptual disambiguation, in which the folk psychological concepts are refined and made fit for purpose, can alleviate these problems. I will first introduce the idea of what I call 'cognitive scientific discourse', i.e. the discourse within which cognitive science offers explanations and predictions, and contrast this with the folk psychological discourse that was the topic of the previous three chapters. I will then consider four cases where the two discourses seem to come into conflict, and where I think either refining or abandoning folk psychological concepts will help us to make progress. In doing so I will draw on the positive account of folk psychology that I have presented in the first half of this thesis, which will allow me to argue that in each of these cases it is not simply that folk psychology has given us a *false* picture of how the mind works, but rather that the concepts it provides should never have been applied to fine-grained cognitive scientific descriptions in the first place.

The four cases that I will consider are: 1) the false belief task in social cognition, 2) the scientific taxonomisation of sensory modalities, 3) the extended cognition debate, and 4) the emerging predictive processing paradigm. Each case study is individually intended to illustrate a particular example of how folk psychological concepts might fail us, and taken together they give a general sense of the kinds of problems that might arise. My aim in this chapter is not to conclusively prove that folk psychology is unsuitable as a source of scientific concepts, but rather to give an initial indication why I think the technical use of folk psychological concepts might be problematic. In each case I will develop a version of what I call the disambiguation strategy, which revolves around either distinguishing different senses in which a folk psychological concept might be applied, or else coming up with new concepts that better capture the phenomenon being described.

In the next chapter I will consider more generally the status of folk psychological concepts as natural kinds, and argue that folk psychological kinds do not typically qualify as cognitive scientific kinds. Finally, in chapter 6 I will examine recent debates about cognitive ontology, and propose a novel methodology for developing new cognitive scientific concepts, whilst retaining a positive role for folk psychology as an initial source of basic, unrefined concepts.

What do I mean by cognitive scientific discourse? Chapter 1 introduced the idea of folk psychological discourse, which I used to refer to the broad set of folk knowledge and social practices that people deploy every day to understand and explain each other's behaviours. My intention in referring to folk psychology as a 'discourse' was to capture the sense in which it goes far beyond mental state attribution, and includes further activities and practices such as behavioural predictions, narrative competency, and socio-normative regulation. Similarly, by referring to a cognitive scientific 'discourse', I want to indicate that the use of folk concepts in cognitive science goes beyond the explicit use of terms such as 'belief' and 'desire' in academic articles, but also includes more subtle effects such as the influence that a researcher's own folk psychological intuitions could have over theoretical and experimental design, or the lingering implications that a word drawn from the folk psychological lexicon carries, even after it has been given a rigorous and technical definition. So my analysis of folk concepts in cognitive scientific discourse will require a certain amount of reading between the lines, so to speak, as there may still be an influence of some kind even when scientists seem to be aware of the inadequacies of folk psychological concepts and are careful with the terms that they use.

I should also comment briefly on the level of detail (or lack of it) that will be necessary for these case studies. Where relevant I will describe the technical details of a theory or experiment, but in general my aim is not to get bogged down in specifics, but rather to give a general sense of the kinds of issues that I see arising when folk psychological concepts are misapplied in cognitive scientific discourse. In some cases it may well turn out that the use of a certain folk concept was more

suitable than I first imagined, but I am confident that the general structure of the problems that I identify in this chapter will continue to apply in the majority of cases, including cases that I don't discuss here.

4.1 – The false belief task and the puzzle of retrogressive development

The false belief task in social cognition was initially developed to test when infants (and non-human primates) first become aware that other people have minds. Initially it was thought that this capacity, generally known as a 'theory of other minds', first appears around the age of four or five, but recent research pioneered by Baillargeon and colleagues has demonstrated that children at least as young as 15 months can pass a non-verbal version of the task. In this section I will present both verbal and non-verbal false belief tasks, and consider some attempts to explain the apparently puzzling development of false belief attributions, before arguing that one way of resolving this puzzle is to simply disambiguate between (at least) two senses of the folk psychological 'belief' concept.

4.1.1 – Verbal and non-verbal false belief tasks

The original false belief task was conducted by Wimmer & Perner (1983), and later developed by Baron-Cohen, Leslie, & Frith (1985), whose version is best known today. The experiment consisted of a child being presented with a pair of dolls ('Sally' and 'Anne') playing with some toys. Sally places her toy in a box and then leaves the scene. Whilst she is gone, Anne moves the toy and hides it in another box. Then Sally returns, and the experimenter asks the child where Sally will look for her toy. The correct answer is that she will look where she originally hid it, but prior to around the age of 4 children tend to answer incorrectly, stating that Sally will look where the toy *actually* is, rather than where she last saw it. The received interpretation of this experiment is that it demonstrates that prior to acquiring a theory of mind at age 4, children are unable to attribute false beliefs to others. The experiment has been replicated many times since, and in many different formats, with fairly consistent results. (See sections 1.2.1 and 2.2.1 for further discussion.)

A 2005 study conducted by Onisihi & Baillargeon presented infants with a non-verbal version of the false belief task, and was able to elicit apparently positive results from children as young as 15 months. The non-verbal false belief task makes use of the violation-of-expectation method (see e.g. Baillargeon 2004), which uses infant looking time as a proxy for what they expect to see. Infants tend to look longer at events that surprise them, and so we can use surprise as an indicator of what the infant expected (or did not expect) to see. In the study conducted by Onishi & Baillargeon infants were first familiarised with a scene where an actor moves a toy watermelon between two differently coloured boxes. Once this scene, and variations on it, was no longer of interest to the infant, the actor hid the watermelon in one box, and then disappeared behind a screen. Whilst they were hidden the melon moved from one box to the other, and then the experimenter returned and looked for the melon in either the box where they had hidden it, or the box where it actually was.[18] If the infant is consistently surprised in the condition where the actor searches in a way that contradicts their apparent belief, then the infant is said to pass the test, as they seem to demonstrate some awareness of the actor's beliefs (both true and false). Infants appear to be able to pass this non-verbal version of the false belief task by at least the age of 15 months. Several variations on this study have since been performed, and the age at which infants can pass versions of this test is has been pushed lower (see e.g. Southgate *et al* 2007, Surian et al 2007, Southgate & Vernetti 2014).

What is strange about these two studies is that children appear to be sensitive to false beliefs at the age of 15 months, but then lose this sensitivity once they begin to talk, until they regain it at around the age of 4 or 5. Retrogressive development of this kind has been observed in other domains; such as number cognition (see Hood 2004 for an overview), but in this case it has yet to be adequately explained. Is it possible that children do actually have a theory of other minds before the age of 4 or

[18] There was also a control condition where the melon had not moved.

5? Or should we instead question whether the non-verbal false belief task is in fact evidence of a theory of mind?

4.1.2 –Accounting for retrogressive development

Since Onishi & Baillargeon (2005) first published their findings, there have been numerous attempts to explain the apparently retrogressive development of theory of mind between 15 months and 5 years of age. Here I will consider three of those attempts, before suggesting in the next sub-section that what at least two of these proposals have in common is a proposal to disambiguate the concept of 'belief'. The first two proposals that I will consider broadly map on to two more general approaches in social cognition (nativist and empiricist), whilst the third posits the existence of two distinct systems for social cognitive processing (and can be given both a nativist and an empiricist interpretation).

The nativist approach to theory of mind holds that we are born with an innate 'mindreading' module that develops along a set trajectory (cf. Carruthers 2006). According to one popular version of this approach the puzzle outlined above is caused by interference resulting from the increased cognitive load required for simultaneous linguistic and social cognitive processing (Carruthers 2013; cf. Newton & de Villiers 2007). The idea here is that one key difference between verbal and non-verbal false belief tasks is that in the former the subject needs to not only keep track of (false) beliefs, but also make sense of the verbal instructions that are given to them by the experimenter. This is particularly difficult for young children who are still learning how to speak, and so until the age of 4 or 5 their attempts to understand the instructions interfere with their ability to pass the task.

The alternative empiricist approach to theory of mind, championed by Gopnik & Wellman (see their 1992), proposes that children learn about the minds of others via an almost scientific process of observation and experimentation. According to this approach the non-verbal false belief task calls for a different kind of capacity than the verbal task, requiring only a basic behavioural understanding of where an object last was in relation to a person, rather than a fully mentalistic

understanding of their beliefs and desires (Wellman 2014). They would therefore deny that pre-verbal infants are in fact able to pass a version of the false-belief task, instead describing the non-verbal task as evidence of a more basic behavioural capacity. The apparently retrogressive development is thus explained by differentiating between the attribution of false beliefs and the tracking of behavioural regularities.

A more recent approach that has been developed partly in response to the non-verbal false belief task is Apperly & Butterfill's 'two systems' theory (2009; Butterfill & Apperly 2013). This is similar to the empiricist approach in that it claims that the two tasks are solved in different ways, but it goes further by positing the existence of two distinct systems that are involved in social cognition. The first of these ('system 1') tracks what they call "registrations" and "encounterings", i.e. subdoxastic states that are still sufficient to pass the non-verbal false belief task, whilst the second ('system 2') operates more like the full blown theory of mind posited by traditional accounts. This two systems account is able to account for the apparent puzzle caused by Onishi & Baillergeon's findings by claiming that system 2, which is required in order to solve the *verbal* false belief task, does not come online until the age of 4 or 5. Prior to this infants must rely on their system 1 capacities, which they lack explicit access to, therefore rendering them unable to give verbal reports about false beliefs.

It is worth noting that in recent research all of the above approaches have come close to some sort of reconciliation. Both the nativist and the empiricist approaches accept that some combination of inherited and acquired capacities are involved in social cognition, and the two systems account looks, from at least some angles, very much like a more fine-grained version of the modular nativist account. A unified account of the false belief task might incorporate aspects of all three approaches by positing two modular systems that become active at different stages of development, perhaps requiring empirical stimulus in order to develop properly. Probably none of the original theorists would be entirely happy with this

compromise, but it could allow for an explanation of the experimental data that is at least somewhat acceptable from every theoretical perspective.

4.1.3 – Disambiguating 'belief'

Here I am not so much interested in which of these accounts is correct, but rather what they have in common, which is, at least in the case of the second two, a proposal to disambiguate our folk concept of 'belief'. Both the empiricist characterisation of a basic understanding of where someone will look and the two-systems theory of 'registrations' and 'encounterings' do away with the attribution of belief as a prerequisite for passing the non-verbal false belief task. A belief, understood in something like the folk sense, is a fairly rich concept which includes the potential for inferential reasoning, whilst a registration simply indicates knowledge of where an object was last seen, and carries no further conceptual baggage. Distinguishing between beliefs and some more basic kind of attribution opens the door to a fragmentation of the folk psychological concept of 'belief', which turns out to simply be too coarse grained to capture the distinction between verbal vs. non-verbal false belief tasks. Alternatively, we could say that solving this problem requires coming up with an entirely new concept, rather than distinguishing between different kinds of belief, but the point remains the same: where once we referred to both cases in the same way, we must now disambiguate between two distinct concepts.

Carruthers, on the other hand, continues to insist that Onishi & Baillergeon's study simply demonstrates that 15-month old infants *must* have an understanding of belief (Carruthers 2009, forthcoming).[19] Carruthers' position, according to Hutto (2016: 8), "risks systematically confounding descriptions of what is being done with substantial accounts of how agents manage to do what they do". Whilst both verbal

[19] Although it is important to note that Carruthers has a notoriously thin notion of belief, one that is perhaps closer to Apperly & Butterfill's 'registration', potentially making his insistence on this point somewhat more understandable. Nonetheless, his continual usage of the folk term 'belief' to refer to whatever is picked out by this thin notion is likely to invoke misleading associations, especially as he seems to think that this thin notion will eventually develop into a full blown concept of belief.

and pre-verbal infants seem to be able to accomplish versions of the same task, we must not assume that the way in which they accomplish these tasks is also the same. Attaching the label 'belief' to both cases makes it harder to question this assumption.

The issue here seems to be the way in which the rich folk notion of belief has been used to define the problem space. Characterising the non-verbal task as a species of false belief task, rather than as a more general social cognition task, already invites researchers to conceive of it as a fairly cognitively demanding task, and implicitly rejects the possibility of an infant 'solving' the task without tracking or attributing false beliefs. Once this conceptual baggage is discarded some space is opened up for reconceiving the task in different ways, such as in terms of registrations and encounterings. This is not to say that any of these alternative conceptions are necessarily correct, but rather that by revising our use of the term 'belief' we can consider theoretical possibilities that might otherwise have been hard to make sense of. The more general point here is that our use of pre-theoretical concepts, such as belief, might sometimes make it hard to imagine other ways of categorising phenomena.

4.2 – Taxonomising sensory modalities

I was brought up being told that people have five senses: sight, smell, taste, hearing, and touch. However, it's not at all clear how well these senses match up to biological reality, or even if this common knowledge is a stable intuition, rather than merely an artefact of our culture or language. In this section I will consider both of these questions, and argue that sensory taxonomisation is another case where our folk intuitions lead us astray. First I will consider the status of the folk taxonomy itself, and suggest that even this is ambiguous once you take into account the way that senses are taxonomised in other cultures. Then I will turn to scientific taxonomisation, and present evidence against a simple distinction between five senses. Finally I will consider two distinct strategies for reconciling our intuitions about the senses with these empirical findings, and suggest that a combination of the two strategies provides the best way to move forward.

4.2.1 – Folk intuitions about the senses

Where does the idea that we have five senses come from? MacPherson (2011) traces it back to Aristotle's *De Anima*, in which the necessity of there being only five senses is argued for on the basis of Aristotle's understanding of the material elements. Nudds (2004: 35) has claimed that it is "obvious" that we have five senses, on the basis that our folk taxonomy is simply not the kind of thing that can be disproven. He goes on to argue that any attempt to give an alternative account of the senses would be simple "changing the subject" (*ibid*). However, this argument only goes through if the folk taxonomy does in fact consist of only five senses, and it is not at all obvious that this is the case.

In chapter 2 I presented evidence for there being cross-cultural variation in folk psychological intuitions, and this evidence extends to folk intuitions about the senses. For instance, the Hausa of Nigeria refer lexically only to *gani* (sight) and *ji*, which is a multimodal sense capturing anything other than sight, as well as emotional understanding and intellectual knowledge (Ritchie 1991: 194). The term *ji* is in fact more dominant in Hausa descriptions of sensory experience, and although context determines to some extent which modality is being referred to, the distinction is certainly less clear-cut than that which we in English make between sound, smell, taste, and touch (not to mention knowing or understanding).

Another, perhaps more intellectually sophisticated example is the inclusion in Buddhist philosophy of a sixth, inward looking sense (Hamilton 2001: 53). This is perhaps comparable to what, in Anglo-American philosophy, is referred to as 'introspection'. Whilst it might seem odd to categorise an internal process as a sense, this is exactly what happens with kinesthesis or proprioception (the sense of where one's own body parts are), which is a common addition to the typical sensory taxonomy (see e.g. Heil 2011: 151).

The details of these examples are not important, but it means that Nudds' assertion that there are "obviously" only five senses cannot be quite right, or at least must be relativized to a specific culture. Nudds does qualify his claim by proposing

that the senses might be social, rather than natural kinds, but even if this is so it remains the case that folk intuitions may not be a good starting point for a scientific study of whatever it is that our sense organs do. I will return to the question of folk kinds and natural kinds in the next chapter, but for now I turn to the scientific study of the senses.

4.2.2 – Scientific taxonomisation of the senses

If anything, the scientific and philosophical taxonomisation of the senses is even more complicated than the folk taxonomy outline above. Macpherson (2011) suggests that taxonomies can be built according to four main sets of criteria: representational format, phenomenal character, sense organ, and proximal stimulus. When it comes to the traditional fives senses these criteria more or less match up; sight has a distinct representational format (2½d image), phenomenal character (visual experience), sense organ (eye), and proximal stimulus (light), as does hearing, etc.[20]

However, research throughout the 20th century has cast doubt on whether this traditional taxonomy genuinely accounts for the full range or complexity of human (and non-human) sensory modalities. Research of this kind can point in one of three directions: either suggesting that two previously distinct modalities should be considered one and the same, or suggesting that an existing modality should sub-divided, or even that a completely new modality should be added.

An example of the first direction is the discovery of the close connections between our senses of taste and smell. If one has a blocked nose it is common to experience a dampened sense of taste, and there are well-understood mechanisms that are responsible for this effect (primarily those involved in retronasal olfaction, see below). Is taste, then, merely a subset of our sense of smell, or it is a distinct sense that is simply very reliant on the sense of smell? The folk taxonomy does not give us any clear answers to this question.

[20] Taste and smell might be an exception here – it is debatable whether or not they genuinely rely on different sense organs, and they both respond to essentially the same proximal stimulus.

This influence goes in the second direction too: contemporary research typically distinguishes *two* senses of smell, one of which occurs in the nose (orthonasal), and the other of which is the result of odorants in the *mouth* (retronasal), and typically contributes to what we commonly describe as taste (Rozin 1982). Should we say therefore that the retronasal sense is a component of taste (despite it having more in common with smell-mechanisms), or should we distinguish three chemical senses, two in the mouth and one in the nose, or should we perhaps just collapse all three into a single chemical sense? It doesn't seem like there is going to be a clear empirical answer here; rather it is a more conceptual question about how we think mechanisms or kinds ought to be individuated. I will return to this question in the next chapter, where I consider the status of folk psychological kinds as natural kinds.

Finally, consider the many potential candidates for additional senses beyond the traditional five. Proprioception or kinesthesia is our sense of where our own body parts are, and is commonly considered an additional sense, distinct from touch. Then there is our sense of balance, governed primarily by the vestibular system in the inner ear, but also influenced by what we can see (if, indeed, we can see). Beyond these there are additional senses that humans may acquire either through practice (echolocation, see Thaler & Goodale 2016) or via the means of sensory substitution devices (magnetic, e.g. Nagel *et al* 2005; infrared, e.g. Thomson *et al* 2013). Sensory substitution devices can also be used to create 'mongrel' senses, for instance by connecting light receptive sensors to a headset, allowing one to 'hear' colours (Montandon 2004: 32-4). How should we taxonomise additional senses of this kind?

As should be clear, there is no real scientific consensus on how the senses should be taxonomised, and not much hope of one being reached anytime soon. The folk psychological taxonomisation, which seems to carve things up differently to contemporary cognitive science, could be seen as no worse than the scientific taxonomy in this regard. If there is no agreed upon scientific taxonomy, then perhaps the folk taxonomy could provide some stability or guidance. In the next section I will consider two strategies to deal with this state of affairs, and suggest that the two

strategies can be fruitfully combined. Using this combined strategy, I will argue that our competing scientific taxonomies can be reconciled once we distinguish between explanations and systems.

4.2.3 – Explanatory pluralism and systemic disambiguation

An increasingly popular response to the apparent mess that is the taxonomy of the senses is to adopt some form of pluralism, wherein there are multiple, non-competing ways of carving up the sensory modalities. Fulkerson (2014) argues that our division of the senses depends on the explanatory project that we are engaged in. For example, the human thermoceptive system consists of receptors in the skin that detect changes in temperature. It contributes to our sense of touch, our proprioceptive/kinaesthetic sense, and also our nocioceptive system. Should it be classed as a separate sense, a sub-system that contributes to these senses, or something else entirely? According to Fulkerson, the answer depends on the explanatory project that we are engaged in. If we are interested in how humans detect distal objects, then we can treat the system as part of our sense of touch, whilst if we are more interested in purely physiological processes, then it is better treated as part of the nocioceptive or thermoregulatory systems.

Insofar as it accurately describes explanatory practice I agree with this account, but I think it misses something important. By breaking down the various ways in which the thermoceptive systems interacts with other sensory systems, Fulkerson has actually successfully disambiguated a distinct sensory subsystem. In effect we can be explanatory pluralists whilst acknowledging the objective existence of distinct subsystems, such as that which Fulkerson describes. The important point here is that we should keep the functional categorisation of physiological systems distinct from the role that they play in both folk psychological and scientific explanations. If our taxonomisation of the senses is aimed primarily at explanatory adequacy, then we should be pluralist about what kind of senses there are. However, if we wish to be more (physiologically) precise, we can disambiguate discrete

subsystems (such as the thermoceptive system) and describe in detail how they interact with one another.

In practice the kinds of explanations given by folk psychology and scientific psychology typically have different aims in mind. For example, whilst for everyday purposes it makes sense to distinguish between taste (proximal) and smell (distal), for scientific purposes we might be more interested in distinguishing between smells that enter through the nose (orthonasal) and smells that enter through the mouth (retronasal). Rather than making a definitive statement about what smell 'actually is', we might want to adopt a pluralist attitude where what we mean by smell depends on the explanatory project that we are currently engaged in. Nonetheless, in both cases we should be able to agree on the underlying physiological mechanisms that are involved in picking up and processing chemical traces from the air. So it is possible to individuate sensory mechanisms without saying anything conclusive about the status of the pre-existing folk psychological concepts. I will return to this question of mechanistic individuation towards the end of chapter 6.

There is a parallel here with the individuation of concrete computational states and processes, which I have elsewhere argued can be achieved by referring only to the physical structures involved (see Dewhurst 2016). This leaves the semantic content of the computational states somewhat indeterminate, but only because such content is a feature of our explanatory practices, not of the system itself. Similarly, the individuation of 'folk' senses such as taste and smell might be best thought of as a feature of folk psychological discourse, not of our sensory apparatus itself. This is not to say that the physiological facts of the matter have no role to play in the folk individuation of sense – on the contrary, they tightly delimit the range of possible explanatory interpretations, but there is nonetheless *some* interpretative work that must take place before a folk psychological explanation can be given.

The empirical and philosophical issues surrounding the taxonomisation of the senses are far more complicated than I have made them out to be here, but hopefully I have been able to give a taste of the issues that arise when we conflate folk

intuitions with scientifically respectable evidence. The strategy that I have outlined here, coupling explanatory pluralism with systemic disambiguation, is one that I will return in later sections, and which will be a central theme of chapter 6

4.3 – Extended functionalism and conceptual disambiguation

Clark & Chalmers initial (1998) presentation of the extended mind hypothesis focused on the case of Otto, a man with Alzheimer's whose extreme reliance on his notebook is taken to be sufficient for cognitive extension. More specifically, they present it as an example of an extended *belief*, as Otto comes to believe that MOMA is on 53rd Street by referring frequently to an entry in his diary. Since the very beginning then, arguments for the extended mind have typically focused on the extension of folk psychologically construed mental states. In this section I will explore the implications of this focus and argue that the main lesson to be learnt from the extended mind hypothesis is that our folk concepts of the mind are simply too imprecise to be useful for technical work in philosophy and cognitive science. Once we move away from folk psychological concepts, many of the disagreements surrounding cognitive extension simply dissolve.

4.3.1 – Summary of HEC

The hypothesis of extended cognition (HEC) rests on an idea most neatly expressed by the so-called parity principle:

> If, as we confront some task, a part of the world functions as a process which *were it done in the head*, we would have no hesitation in recognising as part of the cognitive process, then that part of the world *is* (so we claim) part of the cognitive process. (Clark & Chalmers 1998: 29)

This principle is intended to allow for a neutral assessment of what constitutes a cognitive process – we simply have to imagine that a given external process is taking place inside the head, and decide whether we would consider it genuinely cognitive. If so, then to deny it cognitive status when it is located outside of the head would betray an unwarranted "skin-based prejudice" (Clark 2005: 7).

Clark & Chalmers illustrate the strength of this principle with the case of Otto, a man with Alzheimer's who is heavily reliant on a notebook that he carries with him wherever he goes. He uses this notebook to recall the way to the Museum of Modern Art (MOMA, on 53rd Street) in much the same way as his friend Inga uses her own neural memory. According to the parity principle this means that Otto's notebook constitutes a case of extended cognition, as he forms a belief about the location of MOMA in a way that would count as cognitive if it had gone on inside his head. (This summary ignores some important complications and caveats that I will return to in the next section.)

In later presentations HEC became more explicitly linked with functionalism (Wheeler 2010), and the parity principle was reframed as the complementarity principle, which emphasises the divergent forms that cognitive extension can take (Sutton 2010: 193). Wheeler presents an explicit defence of "extended functionalism", arguing that just as traditional functionalism provides a principled basis for non-neural cognition, it can also be applied in order to provide a principled basis for non-brain bound cognition (2010: 247-9). Functionalism's commitment to the multiple realisability of mental states leads quite naturally to the thought that cognitive processes might be realised in distal instantiations. Coupled with a relatively coarse-grained functional analysis of mental states, we find the hypothesis of extended cognition falls quite naturally out of traditional functionalism. After considering some classic criticisms of HEC, I will turn to a more specific attack on extended functionalism, i.e. the variety of cognitive extension that relies on a functionalist theory of mind.[21]

4.3.2 – Classic criticisms

In this section I will rehearse two of the most famous criticisms of HEC: Adams & Aizawa's "coupling-constitution fallacy" and Rupert's "hypothesis of embedded

[21] A functionalist theory of mind is simply one that defines mental states according to their functions, rather than according to their physical structure or according to some other criteria (cf. Levin 2016).

cognition". I introduce these criticisms here as they highlight how the extended cognition debate has focused on assessing whether or not mental states, *understood in the folk psychological sense,* can be attributed to objects beyond the boundary of brain and body. Later on I will suggest that both criticisms can be dissolved if we move away from coarse-grained folk psychological characterisations of cognition – although in the process cognitive extension, in the classical sense, may also be dissolved.

The coupling-constitution fallacy, according to Adams & Aizawa, is the mistaken inference from the claim that something is tightly coupled with a cognitive system to the claim that it is constitutive of that system (2010: 68; cf. Adams & Aizawa 2001). They argue that when Clark & Chalmers claim that Otto's notebook is partially constitutive of his cognitive system, all we should really conclude is that it is tightly coupled with that system. Adams & Aizawa go on to propose several criteria that might qualify to demarcate cognitive from non-cognitive systems, including intrinsic intentionality (*ibid*: 69-73) and distinct patterns of causal processing (*ibid*: 73-9).

The most common response given to these arguments by defenders of HEC is that Adams & Aizawa are simply begging the question. By proposing a "mark of the cognitive" that excludes HEC, they are guaranteeing that the argument will be over before it has even begun. Instead what is needed is some principled definition of mind or cognition that everyone can agree on. Later on in this section I will suggest that our folk intuitions about the mind are simply too vague for this purpose, as they can easily be construed so as to favour either side of the debate (see Clark & Prinz, ms., for discussion). For example, Adams & Aizawa could claim that the folk concept of the mind is inherently representational, citing intuitions about the 'minds eye', etc., whilst detractors might deny this claim, appealing to more embodied intuitions about the mental.

Rupert takes a different approach when he formulates his alternative hypothesis of *embedded* cognition (HEMC), which can account for everything that HEC does without positing literal cognitive extension (see Rupert 2004: 395-7). The

hypothesis is essentially identical to HEC, except that whenever HEC refers to some feature of the environment being *constitutive* of cognition, HEMC instead refers to that feature as making a causal contribution to internal cognitive processing, in the same way that a temporary buttress might make a causal contribution to the structural integrity of a wall without being constitutive of that wall. Thus HEMC seems able to acknowledge the important contribution that the environment makes to cognitive processing without thereby allowing that the environment itself is part of the cognitive system. Rupert argues that as HEMC is a more conservative hypothesis, we should favour it – all else being equal. Rupert considers two responses, but here I will discuss only the second, which focuses on the role of natural kinds in cognitive scientific explanation.

Clark & Chalmers claim that attributing extended beliefs to Otto helps to pick out something "more akin to a natural kind", which makes the notion of belief more useful in cognitive scientific explanation (1998: 14). They support this claim by appealing to the functional similarities between how Otto and Inga use their respectively external and internal 'memories', arguing that "an opponent has to show that Otto's and Inga's cases differ in some important and relevant respect" (*ibid.*), before going on to deny that there are any such respects.

Rupert identifies this kind of strategy as attempting to demonstrate that HEC "provides the most empirically powerful framework for research in cognitive science" (2004: 407). He argues that for a specific example (memory) this does not seem to be true. Internal and extended memories appear to constitute separate kinds, he argues, as they encode information in different ways and have distinct functional profiles in terms of error rate, recall speed, and so on. Positing a single weakly defined kind, "generic memory", does not seem to serve any explanatory purpose (*ibid*). At this point the argument seems to stall somewhat, due perhaps to some confusion around what exactly cognitive scientific kinds are supposed to look like. In the next chapter I will consider the relationship folk psychological and cognitive scientific kinds in more detail, but for the time being I will move on to the main

argument that I wish to focus on in this chapter, Sprevak's *reductio ad absurdum* of what he calls 'extended functionalism'.

4.3.3 – Sprevak's extended functionalism

We have seen how some of the initial responses to HEC were meant to work. Sprevak (2009) argues that not only are these responses unsuccessful, but also that the common functionalist framework that both proponents and detractors of HEC tend to be committed to in fact entails a radical and unconstrained form of cognitive extension. He takes this to be a reason to reject the classical functionalist account of cognition, but in the remainder of this section I will argue that this apparent entailment actually highlights several weaknesses of the folk psychological conceptual taxonomy. First I will rehearse Sprevak's argument that functionalism entails radical extension.

A primary motivation for functionalism is multiple realisability – the intuition that provided they fulfil the correct functional role, a mental state should be realisable in any physical instantiation (Bickle 2016). This is sometimes expressed as the 'Martian intuition', i.e. the intuition that a silicon-based Martian should be capable of realising the same mental states as carbon-based Earthlings. In order to preserve the Martian intuition our functional definitions of mental states must be coarse-grained enough to allow for some incidental variety in how cognitive systems function (Sprevak 2009: 11-2). For instance, it could turn out that Martians process information in a way that leads them to draw different inferences to us, but we might nonetheless want to attribute beliefs to them.

Sprevak's main claim is that any form of functionalism that is coarse-grained enough to preserve multiple realisability will lead to unconstrained cognitive extension. For any potential case of cognitive extension we can imagine a Martian whose internal cognitive system functions in the same way as the proposed extended system. The parity principle forces us to say either that the system in question is not genuinely cognitive, or that this is a case of extended cognition (Sprevak 2009: 12-3).

For example, in the classic case of Otto and his notebook (see Clark & Chalmers 1998: 33-7), we could imagine a Martian whose internal (semantic) memory consists of fleshy pages and an ink-jet with which it notes down information. It must actively engage an inward-facing eyeball (along with a tentacle to turn the pages) whenever it wishes to retrieve information from this memory store. Both Otto and this Martian are (by stipulation) functionally identical with Inga, who possesses neurotypical human memory. Either we must deny that this Martian has beliefs matching the contents of his internal flesh-book, or admit that Otto has extended beliefs matching the contents of his notebook. Thus, coarse-grained functionalism entails cognitive extension. (Elaborated from Sprevak 2009: 12-13.)

HEC is typically limited to relatively moderate and local cases of extension (such as Otto's diary), but according to Sprevak's argument functionalism entails far more radical cases. Each constraint that has been proposed for HEC is vulnerable to a tailor-made Martian case (Sprevak 2009: 16-20). For example, if we were to exclude Otto's diary on the basis that it is extremely vulnerable to tampering by people other than Otto, then Sprevak could simply propose the existence of a Martian whose (biological, internal) memory contains a user-interface that can be accessed and edited by anyone who wishes. Either we deny that this Martian is truly cognitive, which would breach coarse-grained multiple realisability, or we accept unconstrained extension in the parallel human case. So it seems that functionalism, at least of the coarse-grained variety, unavoidably invites unrestrained cognitive extension.

4.3.4 – Fine-grained functionalism and folk psychological ambiguity

Whilst Sprevak considers the possibility of adopting a finer-grained functionalism in order to avoid radical extension, he argues that any version of functionalism that preserves the Martian intuition will inevitably result in extension (2009: 8). In one sense this might be correct, but I think there is a way of adjusting the grain that at the very least makes for a more palatable conclusion, even if it does not avoid extension entirely.

In fact, the problem is not so much the *grain* of our version of functionalism, but rather the way in which functional roles are assigned to candidate mental states. Sprevak addresses this as well, claiming that empirically motivated "psychofunctionalism", which individuates mental states according to our best understanding of cognitive science, will prove just as vulnerable to radical extension as traditional folk psychological functionalism (2009: 9). I will now try and demonstrate that this is not the case, and that a carefully formulated psychofunctionalism can avoid many of the issues that Sprevak raises. The end result that I am aiming for is that, in most cases, we should be able to leave the everyday folk usage much as it is, and introduce a novel term or concept to more accurately describe the apparently strange cases. Below I present an illustrative example before moving on to the extended cases presented by Sprevak.

A common example of a functionally defined mental state is 'pain', which Sprevak describes as a state that produces anxiety along with typical behavioural responses (2009: 10). This kind of description would normally be thought to accurately capture folk psychological intuitions about pain, but consider the following case: a subject claims to enjoy being pricked by a needle, whilst in all other respects exhibiting the usual physical responses associated with pain (flinching, etc). What kind of mental state should we say that they are in? If pain requires anxiety, then it would seem that they are not in pain, yet they profess quite sincerely to "enjoy pain". It seems that our basic functional definition of pain, as guided by our folk psychological intuitions, is unable to capture this case.

A simple solution is to disambiguate between pain-as-a-physiological-response (sometimes called nocioception) and pain-as-suffering (what we might classically think of as pain). Once we adopt this distinction, it turns out that our subject does not enjoy pain as such (i.e. pain-as-suffering), but rather enjoys the sensations associated with nocioception (i.e. pain-as-a-physiological-repsonse). Similarly, we could imagine a Martian whose physiological responses to harmful stimuli are entirely distinct from out own, yet nonetheless professes to suffer from anxiety when pricked with a needle (cf. Lewis 1980). A functional analysis of our

folk psychological concept 'pain' is simply not precise enough to allow for an adequate description of these kinds of cases, but by adopting a distinction between pain and nocioception, we can better make sense of what is going on. The Martian clearly suffers, and so is in pain, even if it does not undergo anything resembling human nocioception.

Morton (2007) makes a similar point with regard to folk psychology more generally, and advocates focusing on particular capacities and functions in order to get a clearer picture of what is going on in each case. It is this kind of strategy that I think we should pursue in response to Sprevak's argument by asking, in each case, precisely what it is that is extended. In a sense this amounts to tightening the grain setting on our brand of functionalism, but if treated with enough subtlety I think this can be done without undermining multiple realisability. Rather than excluding mentality from physiologically distinct Martians, we will end up clarifying the many different ways in which superficially similar cognitive tasks can be executed. Importantly, it will turn out that the Martians in each case presented by Sprevak have the same capacities as we do, but that these capacities are distinct from the more familiar folk psychological capacities that Sprevak equates them with. The inevitable victim of this process will be our folk psychological intuitions, which break down under the philosophical pressure of the HEC debate.

In order to illustrate how I foresee this strategy playing out, I will run through the three examples of radical extension that Sprevak proposes. Each case hinges on how we interpret an ambiguous folk psychological concept, and by distinguishing between different kinds of extended cognitive process we can constrain the most counterintuitive cases of extension. Note that I am not advocating a change to the folk usage of these terms, but rather the introduction of novel technical distinctions. Such distinctions need only concern practicing cognitive scientists, and philosophers with an interest in cognitive science.

Case 1 – Flesh Pages

We can imagine a Martian born with internal flesh-pages that bear ink-marks encoding innate beliefs. These are functionally identical to the ink-marks in the encyclopaedia that I was given when I was born. As such, we are forced to say that I have innate knowledge of everything written in any book that I have access to, regardless of whether or not I have ever examined them (Sprevak 2009: 20-1).

Here I would suggest that we disambiguate between innate and acquired beliefs (or knowledge, depending on how you want to interpret the situation).[22] Innate beliefs, such as those inscribed upon internal or external ink-marks that someone is either born with or given at birth (I take it that these are functionally identical in all relevant respects), can quite readily be extended, as they require no explicit commitment. Acquired beliefs, however, require deliberate epistemic action, such as choosing to write something down or otherwise learning something new about the world. If an acquired belief is to count as a case of extended cognition it must be deliberately created, excluding the Martian's flesh-pages (and my own external book-pages), but including Otto's notebook. There is still extension in this case, but it is much less radical than Sprevak makes it out to be.

This distinction is superficially similar to the more conventional distinction between occurent and dispositional beliefs (see Schwitzgebel 2015: sec. 2.1). You might think that innate beliefs are simply dispositional, whilst acquired beliefs are occurent, but this is not quite right. Beliefs of both kinds could be either innate or acquired.

Consider Otto's notebook – when he writes down that MOMA is on 53rd street he makes a deliberate epistemic action, and acquires, momentarily, an occurent belief about the location of MOMA. Due to his Alzheimers, he then forgets this almost instantly, but according to HEC he retains an (extended) dispositional belief

[22] Note that this distinction is intended as a novel philosophical characterisation of these cases, not as an elucidation of the folk concept of belief. Nothing in the folk concept makes this distinction, and indeed many folk might be hesitant to admit that the Martian, or even Otto, truly believes anything written either on their flesh-pages or in their notebook. Nonetheless, by making this distinction we can describe in more detail the exact consequences of Sprevak's argument.

about the location of MOMA, which will once again become occurent whenever he looks in his notebook.

Now consider a Martian who is born with a set of internal flesh-pages containing information about the location of MOMA. According to Sprevak's argument this Martian possesses a dispositional belief about the location of MOMA, which will become occurent whenever the Martian chooses to look at those pages. However, what I went to say is that this belief, even after the Martian looks at it, remains importantly distinct from both Otto and Inga's beliefs. Whilst their methods of storage are distinct, both Otto and Inga have deliberately come to acquire this belief, whilst the Martian was just born with it, and until he looks at it will not necessarily have ever explicitly endorsed it. So we can make a distinction between innate and acquired beliefs, and thus limit the *quality* of the cognitive spread that is implied by functionalism, even if we cannot limit the *quantity*.

Case 2 – The Mayan Calendar

We can imagine a Martian born with a Mayan calendar faculty, functionally identical to a program that I can run on my computer. Neither the Martian nor I have ever accessed this faculty/program, but nonetheless we both possess the ability to calculate the precise date of the Mayan calendar, should we ever wish to (Sprevak 2009: 21-2). If the Martian's faculty is considered to be cognitive, then it seems that my own access to this computer program must also be considered cognitive (and thus, constitutes a case of extended cognition).

Here I would suggest that we make a similar move to the previous case, and distinguish between an innate and acquired *faculty*. The Mayan calendar faculty is more akin to a low-level innate faculty like my own circadian rhythm (or a generative grammar), as I was simply born with it and have never explicitly engaged with it in any way. Perhaps it is this kind of faculty, rather than more explicit acquired faculties such as riding a bike or doing algebra, that can be readily extended into artifacts such as computers. Again, this is still somewhat counterintuitive, but less radical than it first appeared.

Case 3 – The Supercomputer

Finally, we can imagine a Martian equipped with an internal supercomputer, allowing him to calculate faster than the most amazing mathematical savant. I could also be equipped with an identical (albeit external) supercomputer, in which case my own (extended) arithmetic would be superior to that of a mathematical savant (Sprevak 2009: 22).

There are two things to say about this case. Firstly, these calculations might be performed in a relatively distinct fashion to the way in which a mathematical savant calculates, allowing us to disambiguate between 'digital' and 'savant' arithmetic (cf. Adams & Aizawa 2010: 75-6). Alternatively, if it turns out that the calculations were performed identically, and that I had sufficiently reliable access to the supercomputer (i.e. sufficient to allow me to outperform the savant in every possible circumstance), we might just have to admit that this is a genuine case of extended cognition. I take it that Sprevak's argument relies on demonstrating the existence of a practically limitless number of 'absurd' extension cases, such that allowing the occasional case through should not by itself rule against functionalism.

4.3.5 – Disambiguating the folk taxonomy

Our folk psychological intuitions about specific kinds of mental states are not the only potential victims of the strategy that I advocated in the previous section. It turns out that the concept of 'mind' itself begins to look increasingly precarious when we take this route, and even the sharp distinction between cognitive and non-cognitive states and processes can be called into question.

Just as folk psychological concepts suffer from a degree of ambiguity, cognition itself is not clearly defined. Despite general agreement with regard to neurotypical human subjects, the precise criteria that a state or process must meet in order to qualify as cognitive vary considerably from theorist to theorist (Hurley 2010: 106). For instance, it is only relatively recently that unconscious processes have been admitted into the domain of the cognitive (Clark & Prinz, ms.), and there

is considerable contemporary disagreement about whether or not "intrinsic intentionality" is required for cognition (see e.g. Adams & Aizawa 2005, Clark 2010).

In each of the above cases it seems plausible to say that *something* is extended, in the sense that we are able to replicate externally a process that the Martian carries out internally, but it is not at all clear that whatever is extended must necessarily be cognitive (cf. Coleman 2011). Couldn't we just as easily have a 'hypothesis of internalised non-cognition', where in each case it turns out that whatever capacity the Martian possesses is not genuinely cognitive? There just doesn't seem to be any pre-theoretical fact of the matter about how we should define cognition, and in the absence of such a definition it seems better perhaps to follow Sprevak in admitting that "[m]ental systems do not form a natural kind" (2009: 29).

There remains a distinction to be made, between extended and non-extended systems, but whether or not we describe these as *cognitive* systems seems to be fairly irrelevant in practice. Consider Otto and his notebook – as Sprevak demonstrates, functionalism seems committed to saying that Otto is part of an external system that is identical to the Martian's internal system. Does Otto have an extended mind? One response is to point out fine-grained functional distinctions between Otto and Inga that appear to rule out this possibility (see Adams & Aizawa 2001: 55-6), but it is quite open to the defender of HEC to reject these as irrelevant to cognition, and there seems to be no clear reason to favour one side or the other. Our intuitive concept of 'mind' might just be another ill-defined folk psychological concept that by itself is unable to adjudicate these debates.

Ross & Ladyman (2010) argue that we should take issues of this kind as evidence that cognitive science is not yet "mature", and that there will be no fact of the matter about HEC until it is. One feature of a mature science, according to Quine (1969), is that it is able to provide accurate definitions of the putative natural kinds that form its domain. In the case of cognitive science, these have classically been assumed to be folk psychological kinds, but what I hope to have suggested here is that folk psychological kinds are simply not up to the job that will be demanded of

them by a mature cognitive science. In the next chapter I will approach this question of natural kinds and folk kinds more rigorously.

In each case presented above the apparent puzzle was resolved by making more precise conceptual distinctions that are not captured by the folk psychological taxonomy. The distinction made in cases 1 and 2, above, is in fact not entirely dissimilar to that which is commonly made between "know-that" (i.e. propositional knowledge) and "know-how" (i.e. procedural knowledge). This is especially apparent in the Mayan calendar case, where I suggested that the capacity that the Martian is born with is something more akin to my innate capacity to keep track of time than it is to my learnt capacity to calculate the date in the North Korean 'Juche' calendar. It is in the former sense that Sprevak *does* possess the capacity to calculate the Mayan calendar date, albeit via his interactions with a techno-cultural artefact. It may simply turn out that 'knowledge' does not unambiguously refer to a discrete set of states and processes, whether extended or not. Of course, it is also always open to us to use folk psychological concepts in a more circumscribed way, rather than rejecting them outright – in a sense this is exactly what I am suggesting we should do by adopting the disambiguation strategy. In practice there is not much difference between distinguishing two different kinds of belief on the one hand, and choosing to use 'belief' to refer to only one kind, and some other term to refer to the other. All that I ask is that we use these terms more precisely, and take care to specify exactly what technical sense we are using them in at any given moment.

4.4 – Alien representations and opaque contents[23]

In this final section I will present and consider recent developments in hierarchical predictive coding, also known as the 'Bayesian brain hypothesis'. This approach, I will argue, is committed to the existence of 'alien' mental states that do not correspond in any direct way with folk psychological attributions. Once again we are faced with an apparent disconnect between scientific and common sense

[23] Much of the material in this section is forthcoming as "Folk Psychology and the Bayesian Brain", in Metzinger & Wiese (eds.), *Philosophy and Predictive Processing.*

understandings of cognition. I will close with a short discussion of how best to reconcile the two, which will serve as an introduction to what is to come in chapters 5 and 6.

4.4.1 – Predictive processing

My discussion will focus on Andy Clark and Jakob Hohwy's presentations of the Bayesian brain hypothesis, which are not entirely identical, but I will note when they differ. I will use the term predictive processing to refer to their versions of this hypothesis. Other versions of the Bayesian brain hypothesis exist (see Spratling 2016 for an overview), and these differ from Clark and Hohwy's in many important ways, but I will not be discussing them here. Below I present a very brief introduction to predictive processing (see Hohwy 2013 or Clark 2016 for a more detailed overview).

Both Hohwy and Clark endorse versions of the Bayesian approach to cognition. Speaking very generally, they claim that the brain's primary function is to generate and test hypotheses about the external world, and that this process is constitutive of cognition. More specifically they focus on a version of the hypothesis known as predictive processing (or predictive coding[24]), which describes a particular way that Bayesian hypothesis testing could be implemented in the brain. I will now present a rough sketch of the cognitive architecture posited by predictive processing, focusing on the details that are most pertinent to my discussion of folk psychology. For a full overview see Hohwy (2013) or Clark (2016).

Predictive processing inverts conventional assumptions about the flow of information in the brain. Rather than starting with raw perceptual inputs that are gradually processed into refined models of the world, it begins with a rich, internally generated model that predicts incoming sensory data. These predictions are then compared with the actual data, and the model is updated accordingly. Overall the

[24] The terms 'predictive processing' and 'predictive coding' are often used more-or-less interchangeably in the philosophical literature, but strictly speaking the latter refers to a specific data compression technique whilst a former refers to the cognitive or computational architecture composed of iterated instances of that technique. Hierarchical predictive coding is thus (roughly) equivalent to predictive processing,

system aims to minimise prediction error, which can be accomplished in two distinct ways. The model can be revised so as to more accurately predict incoming stimuli (passive inference), or the system can act on its environment in order to make its own predictions more accurate (active inference). Which kind of inference is performed (active or passive) will depend on higher-level predictions of the best way to reduce error in the current situation. Thus Clark summarises predictive processing as positing "core perception-attention-action loops in which internal models of the world and their associated precision expectations play key action driving roles" (Clark 2016: 71). By uniting action and perception in this way, predictive processing aims to provide a general account of cognition.

There are a few further features of predictive processing that are especially relevant to my discussion of folk psychology. Predictions can be regarded as more or less precise by the system, with the level of precision being taken into account when updating the predictions. For example, a less precise prediction will be expected to generate some error, and so may not need to be modified too much when an error signal is received. Changes in precision weighting can also drive the system to attend more or less to different sources of stimuli (Clark 2016: chapter 2). Finally, the predictive processing systems described by Hohwy and Clark are hierarchical; they consist of a nested hierarchy of precision/error units, with each level of the hierarchy predicting the current state of the unit below, which is then compared to the actual state of that unit and updated (in the next iteration) in response to any error signals that it receives. This hierarchy bottoms out in units that predict inputs received via sensory transduction, and tops out with a very abstract model, perhaps just predicting general causal laws or regularities. I will describe some further features of predictive processing in more detail as I go on to compare it with folk psychology.

4.4.2 – Predictive processing and propositional attitude psychology

Both Clark and Hohwy have suggested in informal discussion that predictive processing might be incompatible with the folk psychological conception of cognition. Clark has described the content of the predictions as "alien" and

"opaque"[25], and Hohwy has acknowledged the challenge posed by predictive processing to "folk psychological notions of perception, belief, desire, decision (and much more)".[26] Clark has also written that predictive processing "may one day deliver a better understanding even of our own agent-level experience than that afforded by the basic framework of 'folk psychology'" (Clark 2013: 17, repeated in Clark 2016: 82). I take it that what both of them have in mind when they refer to folk psychology is propositional attitude psychology. This is the interpretation of folk psychology that has traditionally been of most interest to philosophers, and as such it is a good place to begin my assessment of predictive processing and folk psychology. In this subsection section I will focus on belief and desire, although the issues raised here will generalize to other propositional attitudes.

Belief

According to the conventional account of propositional attitudes, a belief is a state consisting of a proposition coupled with a positive epistemic attitude, i.e. one that regards it as true, and a belief state will interact with other mental states so as to generate actions in accordance with the state of affairs captured by the proposition being true. It is sometimes said that beliefs have a 'mind-to-world' direction of fit – that is to say, a belief should be modified in response to how the world is, and not vice versa.

On the face of it this kind of mental state seems to fit nicely into the predictive processing story. It is natural to interpret predictions as beliefs about the world, albeit ones that are first generated and then tested, rather than being generated in response to sensory input. Indeed, some researchers have described predictions as beliefs, including Karl Friston (see e.g. Hobson & Friston 2014), Hohwy (2012), and

---

[25] Comment made during BPPA Masterclass on Action-Oriented Predictive Coding, University of Edinburgh, 26th-27th October 2013.

[26] From the comments section of a Brains Blog featured scholar post: http://philosophyofbrains.com/2014/06/22/is-prediction-error-minimization-all-there-is-to-the-mind.asp.
See also (Hohwy 2013: 2) for a description of how he thinks predictive processing might lead us to "radically reconceptualize who we are".

occasionally Clark himself (2016: 129). Given that the term 'belief' is used in a technical sense in Bayesian theory, this might be excusable; however, simply equating predictions with beliefs in the everyday sense would be to ignore a crucial difference between folk psychological beliefs and the predictions invoked by the predictive processing story. The former are usually understood as determinate (you either believe something or you do not),[27] whereas the latter are inherently probabilistic. Rather than simply believing that it is raining, a predictive processing system will assign a level of probability to it raining, and act in accordance with this probability. As Clark puts it,

> Instead of simply representing 'CAT ON MAT', the probabilistic Bayesian brain will encode a conditional probability density function, reflecting the relative probability of this state of affairs (and any somewhat-supported alternatives) given the available information. (Clark 2016: 41)

As a consequence of this, adopting the predictive processing framework will require either an acceptance that the folk psychological concept of belief was never meant to pick out a certain kind of brain state, or an acceptance that (neural) beliefs are in fact probabilistic rather than determinate. I explore the first option elsewhere in this thesis, where I argue that it is a mistake to think that folk psychology has ever been in the business of describing the structure of cognition at the same level of detail as a cognitive scientific theory like predictive processing does. Something like the second option has been explored by Pettigrew (2015), who considers the epistemological implications of adopting a probabilistic notion of belief alongside the more conventional determinate notion. In the next chapter I will also consider the possibility that we should modify our understanding of 'belief' rather than eliminating it from our scientific ontology.

[27] Whilst there is some discussion of 'degrees of belief' within formal epistemology (see Huber & Schmidt-Petri 2016 for an overview), this is quite distinct from how belief is usually understood within mainstream epistemology and philosophy of mind. Pettigrew (2015) discusses some of the implications of adopting a probabilistic notion of belief.

The predictions involved in predictive processing may also be individuated at a much finer level of detail than folk psychological belief attributions usually allow for. Whilst a paradigmatic belief might be about whether or not it is raining, the content of the predictions at some levels of the hierarchy are more likely to be cashed out in terms of fine-grained details of the external world, predicting features such as edges and light gradients rather than the 'middle sized dry goods' that populate the folk ontology. Even at higher levels of the hierarchy, the content of the predictions are still somewhat unusual, as they incorporate multi-modal, emotional, bodily, and other contextual associations. Combined with the probabilistic nature that I described above, the predictions posited by predictive processing begin to look less like the everyday notion of a belief. Clark expresses something like this view himself when he writes that "the looping complexities" involved in predictive processing "will make it hard (perhaps impossible) adequately to capture the contents or the cognitive roles of many key inner states and processes using the terms and vocabularies of ordinary daily speech" (Clark 2016: 292).

However, Hohwy (2013: 60) describes how the relationship between a predictive processing system and a dynamically evolving world could give rise to higher-level regularities that might come to resemble something more like folk psychological contents. He gives the example of perceiving a partially occluded cat, but forming a prediction of a whole cat based on feedback from seeing different parts of this cat at different points in time as it moves behind the occluder. The content of the whole-cat prediction is relatively coarse-grained, but it would in turn predict lower-level perceptions of parts of cats that change over time. The system is thus able to account both for the diachronic appearance of rapidly changing parts of cats, and the more abstract notion of a whole cat who stands behind the occluder and is temporally extended. So, if Hohwy is correct, we might expect to see something resembling folk psychological states at the higher levels of a predictive processing hierarchy, even if these states are different to how we usually conceive of them (i.e., their content is non-linguistic, abstract, and probabilistic, rather than consisting of linguistic propositions with determinate content).

Finally, it is important to recognise the dual nature of predictions. Predictions function both as representations of the world and of ways that the system can act in the world. Via the mechanism of active inference, predictions can be used to motivate and generate actions, a feature that is usually associated more with desires than beliefs. Clark likens this feature of high-level hypotheses to Millikan's (1996) "pushmi-pullyu" representations, which have "both descriptive and imperative content" (Clark 2016: 187). At this point there is a sense in which beliefs, if they were to be identified with predictions, would begin to blur into what we might more naturally characterise as desires. Hohwy himself suggests that perception and belief might both be reconceived as a single notion of expectation (2013: 72), which could go some way towards reconciling predicting processing with propositional attitude psychology, although it would require that we adopt a revisionary approach towards folk psychology. Taken a step further this revisionary approach could also involve collapsing desire in with perception and belief, leaving us with a single kind of mental state that encompasses all aspects of cognitive processing.

Belief, understood as a positive epistemic attitude towards a proposition, does not straightforwardly fit in to the ontological framework of predictive processing. Whilst proponents of predictive processing have occasionally described predictions as beliefs, they have in mind something quite different to the traditional propositional attitude interpretation of folk psychology. Folk psychological beliefs are typically determinate and take linguistic propositions as their argument, whilst predictions are probabilistic and refer to a wide range of distinct contents, most of which are likely to be non-propositional. Nonetheless, it is plausible that we might find something closer to the folk psychological notion of belief at higher levels of the predictive processing hierarchy, where coarse-grained predictions about stable features of the environment are to be found.

Desire

Much like a belief, a desire consists of a proposition coupled with an attitude; only this time the attitude has a world-to-mind direction of fit, and will function

accordingly. If I desire that it is raining, I will not pick up my umbrella (as I would if I believed it was raining), but I might sigh deeply and complain about the heat, or invest in experimental cloud seeding technologies.[28]

As I mentioned above, the predictions posited by Clark and Hohwy's versions of predictive processing are action-oriented. This means that as well as providing a model of the world, they also serve to motivate the system to act via the mechanism of active inference. In this latter capacity they seem to fulfil a role very much like that played by desires in the traditional account of folk psychology. They represent how the system would like the world to be, and coupled with beliefs about the current state of the world, they generate the appropriate actions to bring about this desired state of affairs. Understood in this way we might conclude that it is viable to adopt a mild revisionism, where it turns out that beliefs and desires are both instantiated by a single kind of state, an 'action-oriented prediction'.

However, as Clark draws attention to, there is another sense in which predictive processing seems to do away with desire entirely. Friston, Mattout & Kilner write "crucially, active inference does not invoke any 'desired consequences'" (2011: 157), which Clark interprets as "a world in which value functions, costs, reward signals, and *perhaps even desires* have been replaced by complex interacting expectations that inform perception and entrain action" (Clark 2016: 129, emphasis added). The key issue here is that predictive processing inverts the conventional ordering of action causation assumed by folk psychology. Rather than a desire generating behaviour that leads to an expected outcome, a prediction of an expected outcome is generated first, which then goes on to cause behaviour that brings about that outcome. Desire seems to be relegated to a phenomenal sensation associated

[28] There is another sense of desire that I will not be discussing in any detail in this section. Whilst the kind of desire that I have described here is inherently action involving, there is another kind of desire, which we might call 'existential desire', that might never cause any actions at all. I might desire world peace but believe it to be unattainable, and so not be at all motivated to actually try and bring about world peace. If predictive processing can account for desires of this second sort, I think it will have to be in terms of some much higher level process, taking into account predictions about predictions (i.e. hyper-priors). In this section I focus primarily on those low-level, immediate desires that are thought to be involved in our day-to-day actions, at least according to the traditional propositional attitude account.

with this sequence of events, and does not seem to play any causal or functional role in generating either the behaviour or the outcome.

Clark argues that there need not be any contradiction here. Instead of eliminating desire from our ontology, we can reconceive of it as a consequence of the interaction between predictions and the environment (Clark 2016: 129). Insofar as it allows us to recognise the differences between predictive processing and folk psychology without simply eliminating the latter, I would endorse something like this position, but first I want to mention a further issue that it raises. Reconceiving desire as a consequence rather than a cause of action has the potential for a deeply counterintuitive picture of personal level agency. Rather than being a distinct source of actions, agency (in the guise of active inference) turns out to be nothing more than a tool used by the system to minimise prediction errors.[29] We do not do things because we want to do them; we *feel* like we want to do things *because* doing them will minimise prediction error. As Hohwy puts it, "[w]hat drives action is prediction error minimisation […] rather than what the agent wants to do" (2013: 89). Hohwy presents this as a positive result, unifying perception and action under one single mechanism (Hohwy 2013: 76), but for many this will seem like a sleight of hand, akin to Dennett's attempts to reconcile free will with a deterministic cognitive architecture (see his 1984, 2003). Perhaps this is just a symptom of a mistaken folk conception of agency, or perhaps it points towards confusion between two distinct modes of explanation – either way, the folk concept of desire would turn out not to be doing any significant work in our cognitive scientific explanations of action generation.

The folk psychological concept of desire as an action-motivating attitude is encompassed by the predictive processing notion of an action-oriented prediction, which via the mechanism of active inference is able to act on the world in order to make itself come true. Thus, predictive processing differs from the folk notion of desire in two crucial respects: firstly, beliefs and desires are implemented by a single

[29] Colombo (forthcoming) makes a similar point, in the context of challenging the empirical foundations of the Humean theory of motivation.

kind of state, an action-oriented prediction; secondly, desire is relegated to a secondary status, as it is prediction-error minimization, rather than any personal goals of the system, that drive action. Hohwy presents this as a positive result, offering the possibility of unifying perception and action under one single mechanism. I think instead that what it indicates is that the project of trying to naturalise folk psychology by identifying propositional attitudes with the theoretical posits of our best cognitive science is a mistaken one, as it misconstrues the aim and purpose of folk psychology. In the next section I will consider how a broader interpretation of folk psychology as a folk discourse fits with the predictive processing framework.

4.4.3 – Predictive processing and folk psychological discourse

We can also consider how well the predictive processing story aligns with folk psychological discourse in the general sense that I described in chapter 1. Friston & Frith (2015) have explored behaviour reading and prediction, understood in the context of Bayesian inference. They argue that predicting the behaviour of another requires synchrony between the two brains in question (that of the predictor and the predicted), allowing predictions to be made based entirely on the current state of one's own brain. This sounds somewhat like the simulation theory in social cognition (see e.g. Goldman 2006), which claims that we understand other minds by analogy with our own mind, although Friston & Frith do not make this connection. Given that predictive processing can be interpreted as generating a simulation of the target domain, this similarity is perhaps unsurprising, although there's also a sense in which the heavy emphasis on inference puts predictive processing closer to the theory-theory (see Ravenscroft 2010: sec. 2.1), which posits a literal theory of how minds work as the main mechanism for social cognition. One possibility here is that adopting the predictive processing framework would contribute to the development of a hybrid theory that includes elements of both theory-theory and simulation theory. (See Quadt [forthcoming] for further discussion of predictive processing and social cognition.)

Narratives and social norms both seem to fit very comfortably into the predictive processing framework. Clark writes that individuals may "actively constrain their own behaviours so as to make themselves more easily predictable by other agents" (2016: 286), a suggestion that fits very neatly into the mindshaping account of social cognition presented by Zawidzki (2013). Clark also suggests that personal narratives might "function as high-level elements in the models that structure our own self-predictions, and thus inform our own future actions and choices" (2016: 286). This is very close to the role envisioned for narratives in personal and social cognition by Hutto (2008), and thus entirely consistent with my broader characterisation of folk psychological discourse. Hohwy (2013: 163) describes how difficult even simple behavioural predictions can be, and suggests that a failure to take into account broader contextual features might help explain the social cognitive deficit found in people with autism. So understanding folk psychological discourse within the predictive processing framework might involve telling a story about how high-level predictions of behaviour will involve complex models spanning not only individual agents but also the social and cultural environments that contribute to their behaviour.

The predictive processing account of cognition might even be reliant upon folk psychological narratives and social interaction more generally. Clark has written elsewhere about the importance of niche construction and cognitive scaffolding for human cognition (2008, see section 3.3 of this thesis), and he devotes a chapter of his book on predictive processing to discussing these issues (2016: chapter 8). By conceiving of folk psychological discourse as a form of cognitive scaffolding we can retain an important space for it in our explanations of cognition, even if folk psychological explanations themselves are sometimes hard to reconcile with predictive processing. For example, by helping to regulate human behaviour via the enforcement of social norms, folk psychological discourse might serve as a form of active inference, changing the social environment so as to make it easier to predict. It also provides shared narratives that can help make sense of the behaviour of others, as well as exerting a regulative influence in their own right (see Andrews 2015). So

even if propositional attitude psychology turns out to be a bad model of the cognitive architecture required for predictive processing, it might continue to be pragmatically useful to conceive of people as the kinds of systems that have beliefs and desires. We are then left with the further question of whether the failure of folk psychological discourse to match up precisely to our current best theories of cognition gives us reason to eliminate it, or whether being pragmatically useful is enough to escape this fate. It is these questions that I turn to in the remaining chapters of this thesis.

4.5 – Failures of the Folk Ontology?

In this chapter I have described a number of cases where folk psychological concepts and classifications seemingly fail to capture the complexity of our current best understanding of cognition. In each case I suggested that what is required is either a disambiguation of the folk concepts involved, or in some cases the opposite, when it turns out that folk psychology makes distinctions that are not recognised by cognitive science. In both cases though, what this amounts to is the adoption of a more carefully calibrated grain setting, of whatever sort that is required for detailed philosophical and scientific investigation. In section 4.1 this meant to distinguishing between the attribution of beliefs and the attribution of some more basic state such as a 'registration'. In section 4.2 this meant distinguishing between sensory sub-systems such as orthonasal and retronasal olfaction. In section 4.3 this meant distinguishing between innate and acquired capacities, in order to delimit some of the unrestrained cognitive extension that Sprevak argues is implied by functionalism. Finally, in section 4.4 this meant distinguishing between discrete beliefs and desires, and a probabilistic 'prediction' that seems to do the work of both.

Are these examples of cases where the folk ontology has failed, and should therefore be eliminated? In the following chapters I will argue that this is not quite the case: the folk ontology should be eliminated from *technical scientific usage*, at least in its current form, but this does not mean that it has to be eliminated *from folk usage*. The reason for this is that it was simply a mistake to interpret folk psychology as being in the business of picking out fine-grained cognitive scientific states and

processes, as Fodor and the Churchlands did historically. Once we adopt a broader understanding of folk psychology as a folk discourse, which I presented in the first half of this thesis, we can begin to revise the technical usage of folk psychological concepts without having any (direct) impact on the everyday folk usage. Such technical revisions might themselves have a knock-on effect on everyday usage, in the same way that the popularisation of psychoanalysis had on the folk conception of consciousness (Richards 2000), but this is quite distinct from directly arbitrating folk usage itself.

Chapter 5 – Folk Kinds and Natural Kinds

In the last chapter I presented four case studies that provide some initial evidence for thinking that folk psychological concepts and categories might not capture the fine-grained distinctions necessary for cognitive science, or in some cases the opposite, that they might posit distinctions that cognitive science does not recognise. One way to think about why this might be problematic is in terms of natural kinds – if folk psychology does not pick out natural (cognitive scientific) kinds, then we might have good reason to stop using folk psychological concepts and categories in cognitive science. On the other hand, it is not obvious that cognitive science is itself able to identify genuine natural kinds. If this were the case, then the question of folk psychological kinds would become somewhat moot.

The aim of this chapter is to assess which account of natural kinds, if any, is most suitable for the analysis of concepts in psychology and cognitive science, and subsequently to investigate whether folk psychological concepts pick out natural kinds. It will be argued that even under a fairly permissive account of natural kinds, folk psychological kinds do not typically qualify as cognitive scientific kinds, although they may qualify as kinds of a different sort when applied to whole persons rather than parts of a person. However, even in the latter case folk psychological kinds exhibit looping effects that may mean it is better to describe them as 'human kinds', rather than natural (or scientific) kinds. The upshot of this is that they are not ideally suited to application in (sub-personal) cognitive science.

At a first pass, a natural kind is simply any conceptual category that picks out 'objective' features of the world, whatever those may be. In the most extreme case natural kinds can be contrasted with totally arbitrary categories, such as 'all objects over 10 metres long'. Other than being over 10 metres long, which is stipulated in the definition, these objects lack any unifying features that tie them together. It is this lack of unifying features that makes the category unsuitable for scientific usage – given one instance of an object over 10 metres long we cannot make any reliable

predictions about other objects over 10 metres long (aside from those stemming from their being over 10 metres long).

Another way of putting this is to say that natural kinds must support inductive inference, i.e. we must be able to make somewhat reliable predictions about properties shared by members of the kind. An archetypal natural kind, such as gold, shares most of its properties across different instances, whilst other putative natural kinds such as species (or indeed, mental states) may be somewhat trickier to demarcate.

In section 1 I will briefly review several popular theories of natural kinds and assess how appropriate they are when it comes to kinds in psychology and cognitive science. In section 2 I will consider the relationship between folk kinds and scientific kinds in various different scientific disciplines, and argue that folk kinds are typically (although not always) revised or eliminated as a science matures. Folk psychological kinds in particular have proved more resilient to revision or elimination, perhaps due in part to the looping effects that I describe at the end of section 2. Nonetheless, in many cases they do not meet even the most liberal definition of a natural kind, as I will demonstrate in section 3. In section 4 I will wrap up by considering two further concerns: the implications of adopting a causal theory of reference, and the relationship between type identity theory and natural kinds essentialism.

5.1 – Natural Kinds and Scientific Psychology

Philosophical discussion of natural kinds has a long history, going right back to Aristotle's attempts to systematise taxonomical classification. In the early modern period both Hume and Locke were interested in natural kinds, as were Mill and others in the 19th century. More recently the literature on natural kinds has seen a revival, with a diverse range of topics spanning metaphysics, epistemology, philosophy of language, and philosophy of science. In this chapter I will be focusing mostly on issues in philosophy of science, as I am primarily interested in the role of natural kind terms for actual scientific practice, as opposed to more fundamental metaphysical and semantic issues to do with natural kinds. In this section I will

briefly present some of the more popular theories of natural kinds, and consider how they might be applied to kinds in psychology and cognitive science.

5.1.1 – Essentialist Theories

Essentialist theories posit an 'essence' of some kind as being the defining feature of kind membership. Classically this might have meant a literal essence distinct from the physical qualities of the kind, although modern essentialist theories need not be non-physicalist. Putnam's famous claim that water is necessarily H2O is an essentialist claim (see his 1975), with the essence in this case being identified as the chemical structure of the water molecule. He denies that 'water' could refer to any other chemical molecule, even one with identical macro-structural properties. Kripke (1980) defends a similar position, where the reference of a kind term is fixed by an 'initial baptism', and remains fixed regardless of any new discoveries about the kind. This can lead to counterintuitive implications, such as the referent of a kind eventually turning out to have none of the properties we initially thought it did, but it also allows for stability of reference across scientific theory change. At the end of this chapter I will discuss one implication that causal theories of reference might have for my analysis of folk psychology. This is that rather than revising our scientific usage of folk terms, we should instead interpret contrary scientific evidence as a demonstration that the folk kind in question simply had properties that we didn't know it had. Nonetheless I will argue that even if this were the case, there is still good reason to avoid conflating folk psychological and cognitive scientific kinds.

Essentialist theories seem relatively well suited to identifying kinds in physics and chemistry,[30] where we might expect to find stable sets of properties picked out by microstructural essences. It is less obvious that this will work in the special sciences, such as biology, where categories often lack stable or discrete definitions.[31] Nonetheless, a category such as 'dog' does seem to be treated as a

[30] Although see Hendry (2006) for some doubts about the status of natural kinds in chemistry.
[31] Although see Devitt (2008) for a defence of a form of biological essentialism.

natural kind by biologists, in the sense that one can use species membership as a tool for predicting the likely characteristics and behaviours of an organism.

Similar concerns might apply to psychological kinds, which seem unlikely to exhibit the kind of unique microstructural properties that we find (perhaps) in physical and chemical kinds, or indeed any other unique, stable property that might serve to demarcate kind membership. For this reason it seems unlikely that an essentialist theory is going to be suitable for identifying kinds in psychology and cognitive science. (Although, as I will discuss at the end of this chapter, type identity theories could perhaps be construed as essentialist.)

The theories described above are philosophical theories about natural kinds, but there is another kind of essentialism that is worth mentioning in the context of this thesis: psychological essentialism. This is the (perhaps surprising) fact that children seem to be naturally drawn towards an intuitive version of essentialism when they begin conceptually categorising features of the world (see Gelman 2005 for an overview). For example, children tend to make an early distinction between living and non-living things, attributing internal causes of behaviour to the former but not the latter (even when both are in fact inanimate, as in the case of stuffed toys). Research in this area suggests that the intuitive appeal of essentialism might continue into adult life, perhaps explaining some features of folk taxonomies (see e.g. Ahn *et al* 2001). Thus there is a tension between the descriptive project of identifying those categories that, on a day-to-day basis, we typically think of as natural, such as the categories of folk psychology or folk biology, and the prescriptive project of determining which, if any, of these categories are actually natural. This tension is especially problematic in psychology, where the kinds that we are studying might themselves be shaped by everyday acts of classification (see 5.2.4).

5.1.2 – Cluster Theories

A more promising candidate for psychological and cognitive scientific kinds are cluster theories, which identify kinds with stable clusters of properties. To qualify for

kind membership an individual need only possess a sufficient number of these properties, allowing for variation in the members of a kind. According to these theories membership of kinds such as species depend on possessing some, but not all, of the properties associated with that kind. For example, even though the kind *wolf* might be thought to pick out furry, carnivorous mammal with four legs and a tail, we would still recognise wolves with three legs, or without a tail, or who occasionally ate cheese, as members of the kind. Contrast this with an essentialist kind such as gold, which seems to unambiguously consist in being an object with the atomic number 79 – all and only those objects with atomic number 79 are gold.

Perhaps the most famous cluster theory is Boyd's homeostatic property cluster (HPC) theory, which posits an additional requirement that we must identify a causal mechanism that explains the clustering of the properties associated with a kind (see e.g. Boyd 1991, 1999; Kornblith 1993: 35; Magnus 2012). This is required in order to avoid cases of coincidental clustering, such as the cluster of surface-level properties shared by gold and iron pyrite, and also to allow for the identification of kinds where clustering in insufficient, such as the cacti genus *Opuntia*, whose members include the superficially distinct prickly pears and chollas (see Dupré 1995: 27-8). Other cluster theories include Millikans' (1999) theory of kinds in the special sciences, and Dupré's "promiscuous realism" (1996), described by Cooper (2013: 953) as a "cluster style account".

One feature that most of these account have in common is a recognition of the somewhat arbitrary nature of scientific classification, where what qualifies as a cluster is determined to some extent by the interests of the scientists investigating the phenomenon in question. For example, Dupré describes how the decision to restrict the term 'fish' to non-mammalian aquatic vertebrates had no real basis in the previous (theoretical and non-theoretical) uses of the term, and that in any case the creatures now referred to by the term really have no more in common with each other than they do with whales and dolphins (1995: 29-30). This is not to say that kind membership is entirely observer relative, but rather that there are multiple ways of carving up the world, none of which is necessarily epistemically privileged.

Nonetheless, once a classificatory project has been defined, there will be a fact of matter as to whether, for instance, whales are fish or not (Dupré 1999).

Accounts of this kind seem better suited to the messiness of classification in psychology and cognitive science. Rather than being forced to identify a mental state with an 'essence' (perhaps a distinct neural localisation, see 5.4.2), we can instead give a functional profile along the lines of "state *x* typically causes these behaviours, is caused by these events", and so on. Going one step further, we could also aim to identify a mechanism that gives rise to the functional profile in question, such as a certain pattern of neural activation or, perhaps less strictly, a personal level account of how this state functions. Property cluster theories, therefore, might be a good fit for identifying cognitive scientific kinds.

5.1.3 – Scientific Kinds as Classificatory Programs

In the last sub-section I suggested that property cluster theories might be a good match for kinds in psychology and cognitive science. However, a recent critique by Ereshefsky & Reydon (2015) raises some doubts in this regard. They argue that Boyd's homeostatic property cluster theory actually fails to capture some scientific kinds, such as non-causal kinds, functional kinds, and heterostatic kinds. The second worry, that HPC might not adequately capture functional kinds, is especially concerning when it comes to kinds in psychology and cognitive science, which are often defined functionally. One reason why functional kinds might not qualify as natural kinds according to HPC is that they may not necessarily be caused by the same underlying mechanism. So whilst two instances of the putative functional kind might share many (or even all) of the properties in the cluster, the homeostatic mechanism responsible for the clustering might be distinct, and thus they would not qualify.

Consider a classic case from philosophy of mind: if pain is defined as aversion to a noxious stimuli, then we can imagine a situation where two organisms exhibit all the behaviours associated with pain, despite having radically different physiologies and neural architectures (cf. the discussion of Martian pain in section

4.3.4). Whilst we seem to have both intuitive and perhaps even scientific reasons for wanting to call these two behaviours instances of the same kind, under HPC this would be a case of incidental clustering without an underlying kind. A defender of HPC might wish to bite the bullet, and simply deny that these are two cases of the same phenomenon, but doing so might require rejecting a functionalist theory of mental states. I will return to this topic later in the chapter, when I discuss the relationship between essentialism and type identity theory (5.4.2).

Ereshefsky & Reydon's proposed alternative to HPC is to conceive of natural kinds as being situated within "classificatory programs" that provide a framework within which to demarcate kinds. A classificatory program consists of sorting principles, which categorises kinds within that framework; motivating principles which justify the choice of sorting principles; and the set of classifications that the sorting principles allow. They illustrate this with the example of the "Biological Species Concept":

> Its sorting principles tell us to sort organisms of populations that interbreed into the same species, to sort organisms of populations that do not interbreed into different species, and to sort organisms that reproduce asexually into no species. The motivating principle for the Biological Species Concept is the hypothesis that interbreeding and the existence of relatively closed gene pools cause the existence of stable and distinct evolutionary groups of organisms. (Ereshefsky & Reydon 2015: 979)

This can be contrasted with the "Phylo-Phenetic Species Concept", which employs different sorting principles such as "genetic markers, overall genetic similarity, and phenetic traits", and thus classifies species differently. The two classificatory programs have distinct motivations, and thus each might be better suited in different contexts and circumstances. There is no right answer about which to employ, say Ereshefsky & Reydon, aside from the more general scientific virtues of the theory within which they're embedded. Once you're within a program, however, there will be clear rules, given by the sorting principles, for deciding which species a token organism falls into.

In the context of psychology and cognitive science, it might be useful to distinguish programs such as direct localization, which sort neural regions according to spatial proximity and fMRI activation, from programs that emphasize neural reuse, which focus more on functional connectivity across the whole brain. I will return to this topic in the next chapter, but for the time being will simply note that regardless of how successful it as a general theory of natural kinds, Ereshefsky & Reydon's notion of classificatory programs seems like it might have some application to current debates about neural individuation.

5.1.4 – Pragmatic Theories

With both the property cluster theories and Ereshefsky & Reydon proposed alternative we have seen what might be called a 'pragmatic turn' in the philosophy of natural kinds. Ereshefsky & Reydon position themselves as charting "a middle course between a purely descriptive and an overly normative account of kinds" (2015). Classical accounts have aimed to circumscribe how science *should* carve up the world, whilst more recently there has been a turn towards simply describing how scientists *do in fact* carve up the world. As I hinted at earlier, my own interests fall more towards the descriptive, although there is a normative element in that I think some ways of carving up the world might be more useful to cognitive science than others.

I will now consider some explicitly pragmatic accounts that fall more towards the descriptive end of the spectrum suggested by Ereshefsky & Reydon. These accounts generally aim to identify the role natural kind terms actually play in scientific practice, and to demarcate kinds in line with that practice, rather than trying to prescribe how scientists should use kind terms. Examples include Wikforss (2010), who argues that natural kind terms are only special insofar as they support inductive generalisation, and do not have any privileged status over and above this; Brigandt (2011), who argues that the epistemic interests of scientists should come first when demarcating natural kinds and scientific concepts; Khalidi (2013), who argues for what he calls a "thoroughly naturalistic approach" to kind demarcation,

based on the classifications used by scientists themselves; and most recently Slater (2015), who argues for a "stable property cluster theory" that seeks only to identify which kinds best support our inferential practices, without giving any conclusive answer as to whether they are 'natural' or not.[32]

What all of these accounts have in common is that they place the interests and expertise of scientific communities above those of metaphysicians, and reject any *a priori* theorising about the status of natural kinds. One consequence of this commitment is that if a theory of natural kinds rules out the putative kinds of any particular science, this is a problem for the theory, not the science (Slater 2015 makes this point explicit). So in the context of psychology and cognitive science, a pragmatic approach would first ask which kinds the scientists involved are working with, and then build a classificatory framework around that. Unfortunately, given the wide-ranging disagreement in the contemporary sciences of the mind, simply reading off a unified taxonomy is not really possible. Should we focus on fMRI voxels as the basic functional units, or connectivity patterns, or abstract mental states, or even single neurons? Compare this with biology, where even if there is some disagreement about how to classify species, there is at least agreement that a species concept of some sort is a useful kind to be discussing. Perhaps this points to psychology and cognitive science being immature, or disunified, or possibly even non-scientific, but whichever the case it makes applying a purely pragmatic account of psychological kinds somewhat tricky.

Nonetheless, we can look at the role played by various putative kind terms in more localised psychological theories or debates, and at the very least ask whether these correspond in any useful fashion to the kinds given by folk psychology. In the next section I will consider the general relationship between folk kinds and scientific kinds, before turning in section 3 to several case studies of apparent divergence between folk psychological kinds and scientific psychological kinds.

[32] Kendig (2015) gives a more recent overview of accounts of this kind.

5.2 – Folk Kinds and Scientific Kinds

It is generally acknowledged that scientific taxonomisation frequently begins with a folk ontology of some sort, before gradually developing a more refined scientific ontology. Ereshefsky & Reydon, for instance, state that they "assume that the kinds of science have been and are epistemically superior, on the average, to those posited by ordinary language or intuition (except in cases where ordinary kinds are found to be scientific ones)" (2015), which they do not take to be a controversial assumption (see also Khalidi 2013: 59). However, the relationship between a mature scientific ontology and related folk ontologies is not usually one of reduction or elimination, as has sometimes been assumed in the philosophy of cognitive science. Whilst there has sometimes been an assumption that mental states such as belief and desire, drawn from the folk ontology, ought to either be reducible to physical states of the cognitive scientific ontology, or else eliminated, we typically see no such assumption elsewhere in scientific discourse. In this section I will first illustrate this point with some examples from other scientific disciplines, before looking at the role played by folk kinds in psychology and cognitive science, and finally arguing that rather than attempting to either reduce or eliminate folk kinds, we should instead re-characterise them as 'human kinds' that describe characteristics of whole persons, and are brought into existence via the looping effects of folk psychological discourse.

5.2.1 – Folk Kinds in Physics and Chemistry

Of all the scientific disciplines, physics seems to have most thoroughly transcended the folk discourse. Many of the discoveries of 20th century physics are wildly counterintuitive, and even innocuous terms such as "weight" have technical definitions that are distinct from their everyday usage. However, despite Paul Churchland's best efforts to the contrary (see his 1979), this has not had much of an impact on our day-to-day perceptions and descriptions of the world, which carry on much as though nothing has changed since the discovery that the world is round. On the other hand, it seems safe to say that when physicists refer to 'strange' and 'charm' quarks, they are not risking any conflation with the folk use of those terms.

Physics, as what we might call a fully matured science, is almost entirely divorced from its associated folk discourse.

In chemistry the story is somewhat more complex. In many cases kinds drawn from prescientific folk chemistry, such as "gold" and "water", have turned out to be fairly respectable scientific kinds. Perhaps the relatively transparent nature of the key properties of (macroscopic) chemical kinds has made them more straightforward to correctly categorise prior to scientific investigation. Just by picking up a lump of gold, or examining a sample of water, it is possible to learn a lot about the kind in question. This is not to say that chemistry has not moved beyond folk kinds (consider artificial elements which do not occur naturally on our planet), or that folk chemistry has always (or even often) been correct. The four elements of classical 'chemistry' turned out not to correspond at all to any natural categories, and they are now no longer recognised as fundamental elements in folk discourse.

An interesting case here is that of 'jade', which actually refers to two distinct compounds, jadeite and nephrite (Putnam 1975: 241). Hacking (2007) recounts the history of the term jade, highlighting the extremely contingent nature of kind reference, and uses it to argue against the causal theory of reference defended by Putnam and Kripke. The important point for my discussion is that this is a case where both the folk and scientific usage of kind term has been flexible and contextual, and where there is no clear right answer about whether jade should refer to jadeite, nephrite, or both. Hochstein (2016a) suggests that we should just accept that the jadeite/nephrite distinction matters more to some disciplines (such as geology) than it does to others (such as archaeology). Something similar may apply to the kinds picked out by folk psychological terminology, which are tightly coupled with the sorts of social, economic, and political interests that Hacking discusses in the case of jade (see 5.2.4).

5.2.2 – Folk Kinds in Biology

Biological kinds, as I mentioned earlier, are somewhat messier than physical and chemical kinds, and the relationship between folk biology and scientific biology is similarly messier. Consider a well-known example of an apparently misguided folk kind: 'fish', including gilled fish, whales and dolphins, shellfish, along with anything else whose natural habitat is the sea. The standard story here is that whilst it was understandable for the folk to classify all of these creatures as one kind, it turns out that they were wrong to do so, as whales and dolphins are really mammals, not fish (*mutatis mutandis* for shellfish, etc.). However, as Dupré (1999) has convincingly argued, there is not really any clear scientific definition of fish, so the question of whether or not whales (etc.) are actually fish is something of a moot point, depending more on the whims of any given definition of 'fish' than on any objective matters of fact. The same goes for many other supposedly scientific biological categories that are derived from folk biology, such as 'prickly pear' or 'lily' (see Dupré 1981). His point is not that folk biological taxonomies are useless, but rather that there is frequently a mismatch between what is useful for everyday folk purposes and what is useful for scientific purposes. So when the folk classify all sea creatures as 'fish' they are identifying a useful projectable predicate, insofar as it allows them to make predictions that are relevant to their lives. Skills that allow one to hunt sharks probably apply similarly to hunting dolphins, and it might make sense to specialise as a fish-hunter, even if the creatures that fall under that category do not actually share any common ancestry. The same goes for many other folk biological classifications, and arguing about how or if these correspond to scientific classifications may just miss the point.

Ludwig (2015) has more recently made a similar point about the potential value of paying attention to cultural variation in folk biology, and argues that doing so undermines any simplistic picture of the relationship between folk kinds and scientific kinds. There is both convergence and divergence between folk taxonomies and scientific taxonomies, and focusing exclusively on one or the other risks

ignoring the complex web of factors that determines both kinds of taxonomisation. For instance, folk taxonomies sometimes end up reflecting the needs of the sorter (edible vs. inedible, for instance) over pre-existing natural categories, but in other circumstances can turn out to reveal genuine distinctions that scientific research was previously blind to. Whilst the Itza' Maya classification of bats along with birds (cf. Atran 1998) might seem like an obvious case of a non-natural category, it actually reflects a locally important mechanism for seed-dispersal, which explains the clustering of certain types of plants in certain places in a way that might never have occurred to researchers otherwise (Ludwig 2015: 8). For the Itza' Maya, who are interested in this clustering, it makes sense to put birds and bats in the same category.

Ludwig argues that divergence between folk and scientific categories should not necessarily be taken as a reason to revise either taxonomy – rather the distinct needs of each taxonomical program should be respected, and the knowledge unique to each should be used to aid the other where appropriate. I think this is the right kind of conclusion to draw when we find divergence between folk psychology and scientific psychology, and in the next chapter I will argue that cultural variation in folk psychology can sometimes be a useful tool for revising our cognitive scientific ontology.

5.2.3 – Folk Kinds in Psychology and Cognitive Science

If biology is messy, then psychology and cognitive science are messier still. As my earlier discussions of cultural variation in folk psychology (chapter 2) and the use of folk psychological terms in philosophy and cognitive science (chapter 4) indicates, there seems to be serious divergence both between different folk psychological taxonomies, and between folk psychological taxonomies and the kinds of taxonomies required for scientific psychology and cognitive science. In the next section I will look in more detail at some case studies of divergence between folk kinds and scientific kinds in cognitive science, but for now I will just discuss one illustrative example and consider what conclusions we should draw from it.

The folk kinds belief and desire, along with other propositional attitudes, have been co-opted by traditional philosophy of mind and put to use in constructing a general theory of cognition. According to this theory, cognition consists of translating perceptions into beliefs about the world, cross-referencing those beliefs with the desires of the organism, and generating the appropriate actions to bring about those desires, based on the beliefs currently held (see Fodor 1975 for the paradigmatic version of this theory). Whilst this may accurately reflect a certain folk theory of mind, and may even allow for successful predictions and explanations of behaviour, I have argued in the previous chapter that in many cases it does not match up with the fine-grained details of the human cognitive system. However, rather than taking this as reason to eliminate belief-desire psychology, as the eliminative materialists have traditionally argued, I instead think that we should follow the pluralist approach advocated by Dupré and Ludwig, and concede that folk and scientific taxonomies might just have different pragmatic aims. If one is trying to predict and explain the behaviour of a whole person, embedded in a socio-cultural framework, then the kinds picked out by folk psychology might, in a sense, be perfectly 'natural'. However, if one is wanting to predict and explain the micro-structure of human cognition, then a different set of kinds are needed, cutting the world at a different group of joints. Unless one is committed to an extremely reductionist ontology, it doesn't really make sense to say that either taxonomy is any more objective than the other. If one was committed to an extremely reductionist ontology, then cognitive scientific kinds are unlikely to qualify anyway.

5.2.4 – The Looping Effects of Folk Psychological Kinds

How can it be the case that folk psychology is successful at picking out personal level kinds, but fails to pick out sub-personal kinds? As I argued in chapter 3, one plausible explanation of this discrepancy can be found in the regulative mechanisms described by the likes of McGeer (2007), Zawidzki (2013), and Andrews (2015). These, I will now claim, create the sort of "looping effect" described by Hacking

(1995), thus qualifying folk psychological kinds as "human kinds". A human kind,[33] according to Hacking, is a kind that is responsive to our very act of studying it, and thus one that we cannot accurately describe it without taking into account these 'looping effects'. He gives the example of multiple personality disorder, which prior to the early 1970s was rarely diagnosed, then became more frequent, and is now once again rarely diagnosed. He suggests that this strange phenomenon can be explained by observing that the very act of diagnosing someone with multiple personality disorder may inadvertently cause them to exhibit the symptoms of the disorder, via mechanisms such as patients wanting to live up to their doctor's expectations. Another example is the phenomenon of stereotype conformity, where someone conforms to a stereotype that they either identify with, or are identified with by others (see e.g. Sinclair *et al* 2005; Zanna & Pack 1975). In both cases the act of labelling or categorising someone can cause them to exhibit features associated with that label, features that they did not previously exhibit.

Folk psychological kinds, and the associated phenomenon of folk psychologising, may exhibit the same sort of looping effects. If, as I argued in chapter 3, the very act of attributing a propositional attitude to someone exerts a normative influence for them to conform to that description, then we would see the emergence of a looping effect. Folk psychological categorisation could become a self-fulfilling prophecy, gaining explanatory and predictive success at least partially as a consequence of these looping effects. This would make folk psychological kinds, such as belief and desire, into human kinds like multiple personality disorder. Here it is important to recognise that unlike other human kinds, which are responsive primarily to scientific practice, folk psychological kinds are 'studied' whenever we engage in everyday folk psychologising, and thus will be (potentially) responsive to *any* act of mental state attribution, behavioural prediction, or self-/other- narration (via the mindshaping mechanisms described in chapter 3). Thus the looping effects

[33] Hacking (2000) later suggested the alternative and perhaps more general term "interactive kind", and Khalidi (2013) argues that the looping effects described by Hacking may not be limited to the human domain. I will continue using the term 'human kinds' here as I think it is useful to highlight the particular form of looping effect exhibited by folk psychology.

exhibited by folk psychology are potentially much broader, and much more pervasive, than those exhibited by other human kinds.

Hacking sometimes suggests that human kinds are distinct from natural kinds, but it would be wrong to interpret him as saying that human kinds are not *natural*. Rather he rejects the essentialism implicit in the distinction between natural and non-natural kinds. By distinguishing human kinds from the more usual examples of natural kinds he hopes to draw attention to the unique properties of kinds that exhibit looping effects. Cooper (2004) argues that these unique effects are compatible with human kinds being natural kinds. Insofar as I am willing to adopt a pragmatic approach to natural kinds I am inclined to agree. If a human kind such as multiple personality disorder is able to support reliable inductive inferences, then I see no reason why we should not concede that it is a natural kind (albeit an unusual one, that would not exist were it not for the meddling of psychiatrists). Whilst Hacking rejects the label 'natural' kind, his account can nonetheless be construed as naturalistic, in the sense that he thinks the kinds created by these looping effects are no more or less natural than any other kinds. We could think of this as an extension of the pragmatic accounts discussed in 5.1.4. So, even if folk kinds do not strictly qualify as 'natural kinds' under the traditional definitions, they might still be usefully considered kinds of the looping sort described by Hacking.

5.3 – Case Studies

In the previous section I have suggested that folk psychological kinds, whilst frequently diverging from the kinds deployed in psychology and cognitive science, should nonetheless qualify as 'natural' insofar as Hacking's human kinds qualify as natural. In this section I will consider several case studies that demonstrate not only the unsuitability of folk psychological kinds to scientific practice, but also the looping effects that qualify them as human kinds.

5.3.1 – Concepts

Machery (2005, 2009) has argued that the term 'concept', as a posit of scientific

psychology, does not pick out a natural kind, and thus cannot support inductive inference and should be eliminated from our cognitive scientific taxonomy. His argument revolves around evidence that the higher cognitive processes typically taken to be conceptual "do not constitute a homogenous kind about which many inductive generalizations can be formulated" (2005: 445). Here I will focus on his claim that that "the main psychological theories of concepts have posited three theoretical entities that have little in common, namely, prototypes, exemplars, and theories" (*ibid*: 446). These three theoretical entities are each implemented differently and have a distinct functional profile, leading Machery to conclude that it would be more productive to treat each of them as distinct cognitive scientific kinds. Prototypes identify members of a category in statistical terms, and do not require that a member possesses every property in order to belong to a category (*ibid*: 453-4). Exemplars identify a specific individual who stands in for the category as a whole (*ibid*: 454). Finally, the theory view of concepts says that a concept is a theory about the properties of the members of the class (*ibid*.) Whilst he concedes that the term 'concept' might be used informally to refer collectively to all three, he does not think that this usage has any real scientific application, as the three senses of concept are distinct enough that a claim about concepts in one sense will not necessarily generalise to the other two.

Machery goes on to advocate what he calls a "scientific eliminativism" (2009: chapter 8) with regard to concepts. This kind of eliminativism differs from previous eliminativist strategies in that it focuses exclusively on the scientific usage of the term 'concept', and says nothing about its usage outside of this context. He characterizes scientific eliminativism as aiming to demonstrate that 'concept' does not refer to a natural kind, and hence that "the notion of concept should be eliminated from the theoretical vocabulary of psychology" (*ibid*: 219), if it is to develop into a mature scientific discipline. Eliminativism of this kind is essentially the same as that which I advocate for folk psychological terminology more generally, and as such it will be useful to consider Machery's proposal in more detail.

As described above, Machery first demonstrates that the term 'concept' in

cognitive science seems to refer to several distinct theoretical entities, including (at least) prototypes, exemplars, and theories. Each of these entities, he argues, has little in common with the others, as each has a distinct functional profile and plays a distinct theoretical role. As such, whilst we can use the term 'concept' in an informal sense to refer to this general class of theoretical entities, the class itself does not form a natural kind. Given that natural kinds (of some sort) seem essential to the project of scientific theory building, we should eliminate the term 'concept' from our psychological ontology, in favour of more precise classifications that do in fact pick out (putative) natural kinds. His argument is far more detailed than I have described here (see Machery 2009: secs. 8.2 and 8.3), but all that matters for my purposes is the general pattern: if a scientific term fails to identify a natural kind, we should eliminate that term from our (scientific) ontology in favour of a more precise or accurate term. Importantly, this argument need not say anything whether or in what sense scientific terms are meant to refer (I return to this topic in 5.4.1), but rather appeals on pragmatic grounds to the fact we need something like natural kind terms in order to make scientific generalizations, and so we should abandon those terms that don't lend themselves to making generalizations. So rather than using the term 'concept', we should use the more precise terms 'prototype', 'exemplar', and 'theory' (at least, based on our current scientific understanding, which could change in the future).

Machery does not have much to say about folk psychology, except insofar as folk psychology was the target of the classical eliminativist arguments that he rejects, but I think that the general argument pattern that he develops for scientific eliminativism can in many cases be applied to folk psychological terminology. Essentially, whenever a given folk psychological term fails to adequately capture the complexity of current cognitive science, we should either eliminate or revise our usage of that term in order to more accurately describe the phenomenon in question. This is the strategy that I pursue elsewhere in this chapter, and in this thesis in general.

5.3.2 – Emotions

The received view in the cognitive science of the emotions is that there are distinct emotions with distinct functional and behavioural profiles, a position that Barrett (2006) has described as "the natural-kind view of emotions". This view meshes well with our folk conception of the emotions, and indeed the candidates for 'natural emotion kinds' are typically drawn from the folk ontology. For example, many researchers posit a set of 'basic emotions', typically *anger, fear, sadness, happiness, disgust,* and *surprise* (see e.g. Ekman 1972, 1992).[34] Each basic emotion is supposed to correspond to "a more or less unique signature response (within the body) that is triggered or evoked by a distinct causal mechanism (within the brain)" (Barrett 2006: 30). I will follow Barrett in focusing on anger as a paradigmatic example of putative emotion kind.

Barrett identifies two primary motivations or assumptions underlying the natural kind view of emotions. The first is the idea that each emotion kind "produces a distinct set of responses (a characteristic property cluster)" (Barrett 2006: 30). Anger, for instance, could be characterised as producing a confrontational response (of some sort) to the provoking stimuli, along with a certain set of physiological responses. The second motivation for distinguishing emotion kinds is the discovery of underlying causal mechanisms distinct to each kind. Anger might be associated with a certain pattern of neural activation, allowing for the identification of anger even in the absence of any external responses associated with the first motivation.

Putting the two motivations together, we get a general approach where emotion kinds can be individuated either by a distinct set of responses, or by a distinct causal mechanism, or by some combination of the two. If the emotion kinds identified are labelled using folk psychology terminology, as they typically are, then we have the potential for a scientific vindication of the (apparent) folk psychology of emotions. Barrett challenges this received view on both fronts, arguing that the folk psychological emotions possess neither distinct functional profiles nor distinct causal

[34] Although see Ortony & Turner 1990 for some early opposition to this view.

mechanisms (2006: 33-45). Consider anger: one person might shout and hurl objects when they get angry, whilst another person might become very quiet and still. These behaviours seemingly have nothing in common, and yet we might nonetheless feel confident describing both people as angry. On the other hand, Barrett reports evidence against there being any distinct causal mechanism responsible for an emotion such as anger. Two very similar expressions of anger might be caused by distinct neural mechanisms, undercutting the idea that anger, and other emotions, could constitute natural kinds.

Barrett goes on to propose a constructivist view of emotions where an embodied "core affect" is coupled with a culturally, contextually, and linguistically mediated conceptualisation that leads to the application of a label such as 'anger' or 'fear'. Thus the very same core affect could be interpreted as anger or fear under different circumstances, and distinct core affects could under some circumstances come to share the same label (see Barrett 2012 for more detail). Importantly for my purposes, the resulting position does not deny that folk psychological emotions such as anger are very real phenomenon, but rather conceives of them as partially cultural, rather than purely biological. Emotions kinds could therefore be seen as human kinds, exhibiting some of the same looping effects that we find in the other cases described by Hacking (see 5.2.4).

5.3.3 – Memory

Rupert (2013) has argued that memory, a clear case of a folk psychological kind, does not in fact constitute what he calls a "generic" natural kind (one that is suitable for giving a coarse characterisation of a cognitive phenomenon).[35] This comes in the context of the cognitive extension debate, where he is seeking to undermine what he calls the "natural-kinds argument for the extended mind" (*ibid*: 25). I introduced this argument in the context of the hypothesis of extended cognition in section 4.3, but

[35] Rupert does not specify what account of natural kinds he has in mind, and I take it that, like me, he is primarily concerned with the pragmatics of our choice of scientific terminology, rather than the metaphysics of natural kinds. In fact, he explicitly specifies that he wishes to remain neutral "in respect of the issue of scientific realism" (Rupert 2013: 33fn8).

here I will focus on the implications that Rupert draws for the status of 'memory' as a natural kind.

Essentially, the argument that Rupert is responding to claims that we can identify cognitive scientific kinds, such as memory, that "have a significant number of instances external to the human organism" (2013: 29), and therefore concludes that cognitive science should recognize extended cognitive systems. In response Rupert offers a dilemma: "either the proponent of cognitive extension individuates the relevant causal-explanatory kinds in a fine-grained way or in a coarse-grained (or generic) way" (*ibid*.). If we take the fine-grained route then it will turn out that actual[36] instances of external memory do not share any relevantly interesting properties with the functional architecture of internal memory, whilst in the second case Rupert suggests that the coarse grained kinds we are left with would no longer be of much use to cognitive science. The first horn of the dilemma is similar in structure to the disambiguation strategy that I proposed in the previous chapter, although the conclusion that I drew there was that extended cognitive processes might just be unlike internal ones, in ways that could be of interest to cognitive science. Nonetheless, I agree with Rupert that considering the status of putative cognitive scientific kinds is going to be an important part of the cognitive extension debate.

Focusing in more detail on the second horn, Rupert argues that the sort of generic kinds that are required to support cognitive extension (typically drawn from folk psychology) just aren't really of any interest to cognitive science. He focuses his discussion on 'memory', which was the subject of Clark & Chalmers' original "Otto's notebook" example (see there 1998). The gist of the argument is that whilst memory as a generic kind is a suitable explanandum for cognitive science, there is nothing aside from this explananda that unites the different kinds of memory (declarative, procedural, semantic, etc.), and thus we should reject it as a cognitive

[36] It is important to note that Rupert does not want to deny that cognitive extension is possible in principle, but rather that there are not currently any (or many) actual cases of human cognitive extension.

scientific kind. An analogous situation would be if the generic kind 'weather' turned out to be caused by several totally distinct mechanisms.[37] Here it would seem that the generic kind 'weather' was not doing any useful scientific work, aside from indicating the phenomenon of interest. Rupert's argument is much more detailed, and he discusses several potential counter-examples, but for the sake of space I will not consider them here. His argument along is enough to serve as an example of the kind of reasoning that might problematize the scientific usage of folk psychological kinds.

Despite their apparent failure to pick out cognitive scientific kinds, Rupert reserves a role for folk psychological terms in picking out interesting phenomena, and as a convenient way of referring to "multiple, distinct kinds" that are related in some non-scientific way (such as historically or sociologically) (Rupert 2013: 36). In this respect he agrees with Machery's assessment of the status of the kind 'concept', insofar as Machery has no objection to its continued informal usage. Rupert also indicates a useful role for folk psychology in the initial stages of our formulation of a novel cognitive ontology, a role that I will discuss in more detail in the next chapter.

5.3.4 – Mind and Cognition

Rupert notes that a version of his argument may also apply to 'cognition' as a generic kind (2013: 41-4), a possibility that I hinted at in the previous chapter as a potential outcome of the cognitive extension debate (Clark & Prinz, ms., make a similar suggestion with regard to 'mind'). If cognition only refers to a general class of phenomena that cognitive scientists find interesting, but between which there are no inductive generalisation that can be made, then the term does not really seem to be doing any work as a scientific kind. Say that it turns out that the only characteristic of a cognitive system that anyone can agree on is that such systems are studied by cognitive scientists – in this case there is no *prima facie* reason to expect claims about one such system to generalise to another. Even if we find some slightly more specific criteria, such as that the systems must exhibit flexible responses to

[37] Rupert suggested this example during the Q&A of a talk he gave at "Exploring the Undermind" (University of Edinburgh, July 15th 2016).

environmental stimuli, then it will still be the case that a very wide set of systems are 'cognitive', and that we may not be able to successfully generalise across them.

Rejecting the idea of cognition as a natural kind renders much of the extended cognition debates moot. Perhaps when Otto uses his notebook he is part of an extended system of some kind, but asking whether this is a *cognitive* system may just not be very useful or interesting. I do not think much rests on whether or not cognition is a natural kind, and it seems plausible that, as Irvine (2013) argues for the term 'consciousness', its continued usage is more a reflection of institutional pressures to define a coherent area of study than any real epistemic goals. One upshot of this would be that folk distinctions between what is and isn't a cognitive system probably shouldn't have much of a role to play in determining what the proper domain of cognitive science is (in the same way that folk distinctions between what is and isn't a 'chemical' should not determine the proper domain of chemistry).

Rupert in fact defends the status of cognition as a natural kind, provided that it refers to a particular, well-demarcated concept. His proposal is that we should define cognition "by successfully modelling paradigmatic cases of intelligent behaviour" (Rupert, forthcoming), by which he means human behaviour. Whilst I agree that it is possible to just stipulate a definition like this, I do not think the pre-theoretic usage of the term commits us to any such definition, and we could just as easily have defined cognition in some other way. In any case, what matters for my purposes is the status of the folk concept of mind or cognition, not how we eventually come to use the term in cognitive science.

5.4 – Further Concerns

So far in this chapter I have suggested that cognitive scientific kinds, whatever they turn out to be, are unlikely to correspond neatly to folk psychological kinds. This is by no means a conclusive argument, and I am happy to admit that my conclusion is at least partly hostage to future empirical developments. However, there are also some more theoretical concerns that would apply even if the empirical evidence turned out exactly as I expect it to. I will consider these in the remainder of this

chapter, before turning in chapter 6 to offer a positive account of how we should respond to the discrepancy between cognitive scientific and folk psychological kinds.

5.4.1 – Causal theories of reference

So far in my discussion of folk psychological kinds I have assumed that if there turns out to be nothing corresponding to the functional profile of, say, beliefs in our scientific ontology, then we should stop using the term belief (at least in certain scientific contexts). One response to this assumption, originally made by Lycan (1988) in response to Stich's (1983) eliminativism, is to point out that it relies on a descriptive theory of reference, where if nothing corresponds to the folk description of belief then belief does not exist. If instead we adopt a causal theory of reference, as Putnam and Kripke do in formulating their account of natural kinds, then it turns out that rather than discovering that belief does not exist, we will instead discover that belief is just a very different kind to that which we originally thought it was. This line of argument dominated the debates around eliminative materialism in the 90s, which concluded with Stich renouncing his eliminativism. Whilst my position is somewhat different to that which he defended, I nonetheless need to say something in response to the causal theory of reference.[38]

I should first reiterate that I am not so much concerned with debates about the metaphysics and semantics of natural kind terms, but rather with the pragmatics of their usage in scientific discourse. Even if it were proven conclusively that the reference of folk psychological terms is fixed causally, it would be a further question as to whether it was appropriate or useful to use these terms in cognitive science. In any case my claim is not that beliefs (etc.) don't exist, but rather that they are not the kind of thing that we should expect to find at the fine-grained level of analysis studied by contemporary cognitive science. In this sense I can fully accept a causal theory of reference, but just deny that belief was ever intended to pick out anything other than properties of whole people. This, I argue, is how folk psychological terms

[38] Zoe Drayson and Daniel Burnston both independently brought this issue to my attention at conferences where I presented material from this chapter.

are actually used by the folk, and why it is problematic to try and use them otherwise in cognitive science. Nonetheless, even if it turned out that the correct way of interpreting the semantics of folk psychological concepts was in terms of a causal theory of reference, it would still be the case that the concepts need to be used differently to how they are now.

For example, imagine that we discover that some form of the predictive processing account described in section 4.4 is correct, and that therefore there is only a single kind of mental state, a 'prediction', which carries out all of the functions previously ascribed to both 'belief' and 'desire'. My argument so far has been that in cases such as this we should conclude that there is no space within our cognitive scientific ontology for the folk psychological entities 'belief' and 'desire', but this only goes through insofar as we are committed to a descriptive theory of reference. If we adopt a causal theory of reference, then we could continue using the terms 'belief' and 'desire', but we would have to acknowledge that they now mean (descriptively speaking) something different to what they did originally. Any plausible folk psychological description of 'belief' will surely rule out it being the same kind of thing as 'desire', so if predictive processing were true then it would at the very least turn out that, under the original folk definitions, these terms do not straightforwardly apply to cognitive scientific states and processes.

Rather than eliminating belief, it would turn out that we had discovered something novel and interesting about belief. Nonetheless, we *would* have to start using the term differently, which would mean either re-educating the folk about how to use the term, or accepting that our scientific usage of the term is different from the folk usage. As I have argued in the first half of this thesis, the folk seem to be getting along just fine, so the first option seems to be ruled out. And the second option, i.e. acknowledging that our scientific usage of the term needs to change, is essentially identical to what would happen under the descriptive theory of reference, the only difference being that we could (perhaps confusingly) continue using the term belief. So ultimately not much rests on which theory of reference we decide to use.

5.4.2 – Type identity theory and natural kind essentialism

Thus far I have been assuming a functionalist attitude towards potential kinds in psychology and cognitive science, taking it for granted that if there are any such kinds they will be individuated functionally. Functional individuation is importantly neutral about the physical structure of the states individuated, and therefore contributes to the rejection of an essentialist theory of psychological kinds. If pain were to be individuated according to the kind of neuron involved, for example, then it might turn out that only humans can have pain, a conclusion that mental state functionalism typically wants to avoid. All of this is in line with the current state of the art in philosophy of cognitive science. But what if that was not the case, and it turned out that some form of type identity theory could actually be made to work? Imagine, for instance, that for each putative mental state, whether folk psychological or not, we could identity a unique neural substance that was associated with that state.[39] Whenever a subject was in pain, their c-fibres would fire, and whenever they believed something, a particular d-fibre would fire, and so on. We might then be more inclined to adopt an essentialist theory of psychological kinds, where each kind was identified by its unique neural microstructure.

Nothing too important seems to rest on whether or not we adopt a type-identity theory. Even if it turned out that pain really was just c-fibres firing, and thus only occurred in humans, it would still be the case that cognitive science as a whole would be interested in investigating pain-like behaviours across different animal species. We might need to come up with a new term, such as 'aversion to noxious stimuli' (see 4.3.4), that was neutral with regard to physical structure, but inter-species 'pain' science could still continue under the new name 'noxious stimuli' science.

Indeed, such a result might in fact be more at odds with folk psychology than the functionalist analyses that I have discussed elsewhere, as it would rule out the kind of everyday animal pains that the folk seem quite happy to talk about (as we

[39] Botterill & Carruthers (1999: 39) call this kind of realism about folk psychology 'compositional realism', and note that it is unlikely to be compatible with psychological reality.

would now have to say that non-human animals experience 'aversion to noxious stimuli', rather than 'pain'). In this sense, at least, it seems that the folk concept 'pain' is partially incompatible with (microphysical) type identity theory, as the folk seem to accept that (some) non-human animals may experience pain. Insofar as an essentialist theory of psychological kinds based on type identity theory would go against folk intuitions in this way, my overall point remains much the same – folk psychological kinds are not suitable candidates for cognitive scientific kinds, regardless of whether cognitive scientific kinds are typed essentially or functionally.

5.5 – Folk Kinds as Human Kinds

In this chapter I have argued that even under a fairly permissive account of natural kinds, folk psychological kinds do not typically qualify as cognitive scientific kinds. I began by assessing several different accounts of natural kinds, and suggested that we should adopt a pragmatic account where which kinds count as natural is determined partly by our best scientific understanding of the domain in question. With this in mind, I then proceeded to consider several case studies, each of which suggested that a putative natural kind, drawn from folk psychology, was not suitable for application to cognitive scientific research. Finally I considered two further complications that arise with regard to causal theories of reference and type identity theory, and demonstrated that neither makes a huge difference to my argument.

In the next chapter I will propose a methodology for developing novel cognitive scientific categories in light of the apparent failure of folk psychological kinds. Before moving on, though, I want to return to the positive account of folk psychology that I presented in the first half of this thesis, this time with the aim of maintaining a space for folk psychological kinds outside of cognitive science. In section 5.2.4 I suggested that the mindshaping mechanisms described in chapter 3 might result in folk psychological kinds constituting what Hacking has described as 'human kinds', i.e. kinds that are only brought into existence via human action, and yet nonetheless describe 'real' categories out there in the world. So the category 'belief', for instance, might not exist as a fundamental cognitive scientific kind

(whatever that might mean), but could still refer to a non-arbitrary set of traits and behaviours, and could figure in folk psychological predictions and explanations. Crucially, to say that a kind is socially constructed is not to say that that kind doesn't exist. Barrett, discussing her constructivist theory of the emotions, makes this point very clearly:

> some researchers might believe that arguing against natural kinds of emotion is synonymous with claiming that emotions do not exist. This, of course, is not the case at all. Most of us (at least in this culture) have felt angry and have seen anger in other people. The question is whether anger and other similar emotion categories have an ontological status that can support induction and scientific generalization, and allow for the accumulation of knowledge. (Barrett 2006: 46)

The same goes, or so I want to say, for folk psychology more generally. The fact that folk psychological kinds and concepts are culturally specific and socially constructed does not in any sense mean that they aren't real, but it does means that they are unsuitable candidates for the primitive kinds of a universal cognitive science.

Chapter 6 – Revising Our Cognitive Ontology

In this final chapter I will investigate how best to go about developing a novel conceptual taxonomy for cognitive science, and I will also consider how this novel taxonomy relates to our folk psychological discourse. I will refer to this novel taxonomy as a cognitive ontology, following recent work on ontology revision in cognitive neuroscience. This work will serve as a template for a more general methodology that can be applied across the cognitive sciences, which I will develop in the second half of the chapter. The purpose of this methodology is to provide a unified account of how scientifically adequate terminology can be refined out of an initial foundation provided by folk psychological discourse.

In order to explain the relationship between folk psychological discourse and our revised cognitive ontology, I will draw on the mechanistic account of explanation. I will argue that folk psychological descriptions can sometimes qualify as sketches of mechanisms that require further decomposition before genuine cognitive scientific explanations can be given. The mechanistic account also allows for a non-reductive approach to multilevel integration, which will cast further light on the relationship between the ontologies posited by different branches of cognitive science. This chapter aims to highlight the positive contributions that folk psychology can make to cognitive ontology revision, whilst nonetheless stressing that such revision is necessary, as demonstrated by the case studies presented in chapters 4 and 5.

Section 6.1 will review the current state of the art in cognitive ontology, considering four different approaches to cognitive ontology revision, in order to establish a foundation from which to develop a general. In Section 6.2 I will apply this methodology to several case studies that demonstrate its effectiveness. In Section 6.3 I will consider some general methodological issues and propose a systematic approach to multilevel integration and interdisciplinary convergence in cognitive ontology, based around the adoption of the mechanistic approach to explanation. Finally, in 6.4 I will discuss the role that I think folk psychology can play in the

formulation of novel cognitive ontologies, and outline just how radical the revision to our folk ontology could be.

6.1 – The Cognitive Ontology Debate

Price & Friston (2005) introduce the term 'cognitive ontology', using it to refer to whatever taxonomy of states and processes best captures the functional organisation of the brain.[40] They argue that the traditional ontologies favoured by cognitive scientists have been challenged by neural imaging studies, and propose a systematic revision in order to "facilitate the integration of cognitive and anatomical models" across the cognitive scientific sub-disciplines. Subsequently there have been several distinct approaches to cognitive ontology, including proposals by Poldrack (2006, 2010) and Anderson (2015). This section will review each of those proposals in turn, before considering some potential issues raised by Klein (2012) and McCaffrey (2015).

Machery has recently a presented a useful taxonomy of the various approaches to cognitive ontology (see figure 6.1), characterising them along two axes: the extent to which data from neuroscience is allowed to influence our ontology (from 'not at all' to 'exclusively'), and the extent to which our current ontology should be revised (from 'conservative' to 'revolutionary'). According to this taxonomy Anderson comes out as the most radical overall, favouring a large amount of revision based primarily on neuroimaging data, whilst Fodor serves as an example of the most conservative position possible, wanting to retain our current ontology and dismissive of any potentially subversive neuroimaging data. It should be clarified that Fodor is a conservative in the sense of being highly resistant to changing the folk psychological ontology, not in the sense of (necessarily) reflecting the status quo – in fact it is plausible that many contemporary researchers, both in philosophy and cognitive science, come out somewhere nearer the middle of this

[40] In their 2005 paper they primarily used the term 'functional ontology', but I will follow subsequent usage and use the term 'cognitive ontology'. Nothing much seems to rest on the choice of terminology here.

diagram. In the following sections I will discuss some of these positions in more details, focusing on those towards the revolutionary end of the spectrum.

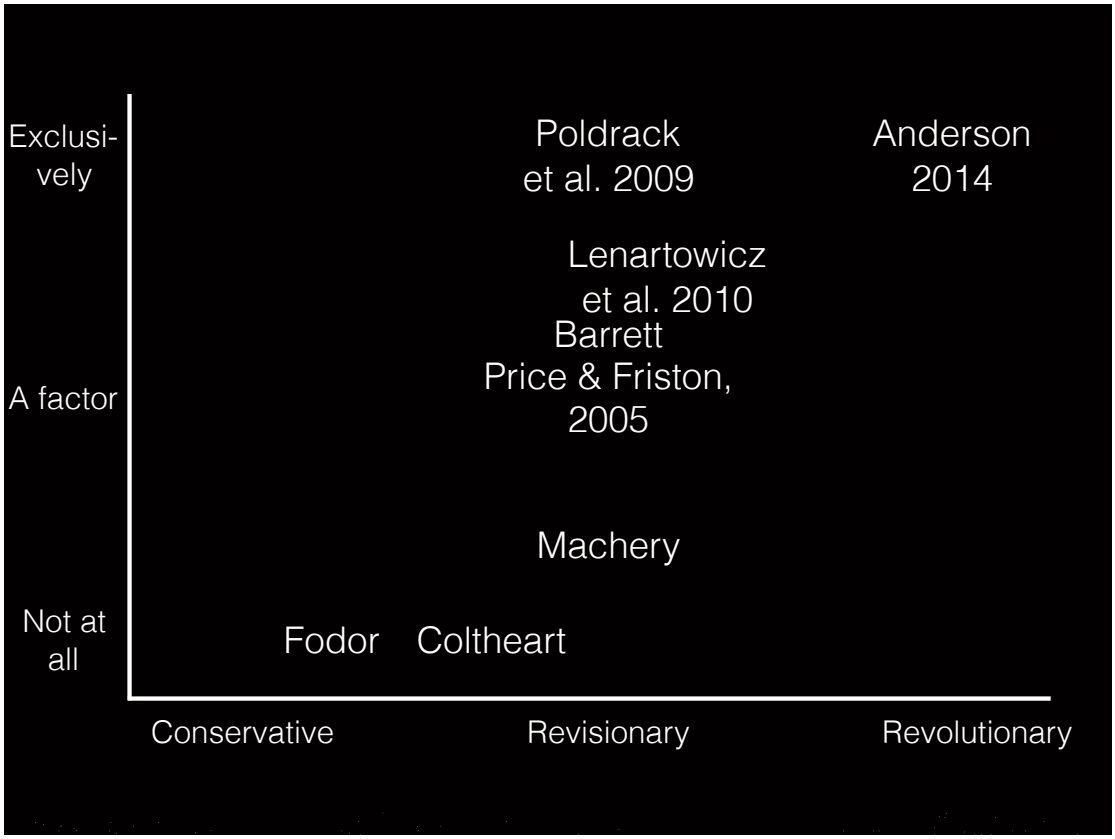

Figure 6.1. Taxonomy of approaches to cognitive ontology, reproduced with permission of Edouard Machery.

6.1.1 – Functional ontologies for cognition (Price & Friston)

Price & Friston's proposed revisions to our cognitive ontology are motivated by the fact that functional neuroimaging has failed to produce a straightforward one-to-one mapping between cognitive processes and anatomical regions of the brain. We see mapping failures of this kind in both directions of fit: cognitive tasks that we take to be discrete "often elicit a distributed pattern of activation" (2005: 262), and discrete brain regions "may be activated by tasks with different cognitive processes" (*ibid.*), precluding the identification of single brain areas with single cognitive functions (as

we currently understand them). They frame this problem in terms of a disjunction between two distinct ways of categorising tasks or processes: a "cognitive set" that is specified according to the behaviours being studied, and an "anatomical set" that is specified according to the brain regions that are activated (*ibid*: 263). It is unclear which of these categories we should prioritise when conducting research, or even if either of them really captures what is going on in the brain. Instead they call for "a systematic definition of structure-function relations whereby structures predict functions and functions predict structures" (*ibid.*). This revised ontology will not conform exactly to either the cognitive set or the anatomical set as we currently understand them, but will rather constitute a novel way of categorising both anatomical brain regions and cognitive functions.

Consider an illustrative example: the left posterior lateral fusiform region is referred to in studies of reading as the visual word form area, because it shows activation when reading written words (Cohen *et al*, 2000). However, it is also referred to in studies of category-specific object processing as the lateral occipital tactile-visual region, where it is associated with the visual attributes of animals (Martin & Chao 2001; Amedi *et al* 2002). The area therefore appears to have at least two functions (visual word processing and animal attribute processing), or perhaps one more abstract function (semantic visual processing?), or alternatively it might just be two distinct areas that overlap significantly. Examining the data in detail, Price & Friston conclude that we should characterise the left posterior lateral fusiform region as performing sensorimotor integration, as this best captures the range of functions that it is associated with. Applied more broadly, this approach would most likely result in a small number of very general functions that accounted for the cognitive phenomena being studied. Whilst folk psychological functions such as 'belief' and 'desire' are similarly general, these new functions would most likely have a more technical/mathematical flavour, as the example of sensorimotor integration suggests.

Price & Friston also consider two associated methodological issues. The first concerns the accuracy of current neuroimaging techniques. Activation within a given

area of the brain could correspond to "spatially proximate but functionally independent neuronal populations" (Price & Friston 2005: 263), further clouding any attempt to put together a reliable cognitive ontology. This problem is compounded by individual differences between subjects, which will further reduce spatial resolution when data are pooled. Apparent failures of our current ontology may turn out to be artefacts of the spatial resolution that we are thinking at, and *mutatis mutandis* for apparent successes.

The second methodological issue that they discuss is that behavioural experiments typically target only a limited cognitive domain at any one time, whilst in fact implicating many different domains. For example, an experiment designed to target word recognition will also implicate general visual processing and attentional mechanisms. This can lead to the activation of neural regions that are only indirectly related to the task, which may then get (potentially) miscategorised as directly related to the behaviour being studied. A task designed around naming pictures of tools may be interpreted as targeting "visual processing, object perception, semantic processing, phonological retrieval, and articulation" (2005: 264), but this ignores the fact that an area associated with "motor processing for hand movements" may also be activated implicitly when categorising tools (*ibid*). The latter area, involved primarily in motor processing, could then be categorised as involved in the former processes, due only to a failure of imagination on the part of the experimenters. The lesson here is that we can never really study a single function in isolation from other cognitive processes, which may activate even if they are incidental to the behaviour being studied. I will discuss the need for a global ontology and convergence across measures in more detail below.

Towards the end of their paper Price & Friston outline several "guiding principles for functional ontologies" (2005: 269-73). They conclude that a good ontology should "have a hierarchical structure that predicts the coactivation of anatomical regions" (*ibid*: 272), where a hierarchical structure is understood as a set of nested ontologies. It should also "enable cognitive processing to be predicted given any distribution of activations" (*ibid*.). Their conception of a revised cognitive

ontology is basically a one-to-one mapping between structures and functions, where neural activation in a given area accurately predicts a single task, and vice versa. These may seem like fairly modest criteria, but in practice they can be very hard to achieve. Furthermore, what constitutes a one-to-one mapping will depend largely on how structures and functions are defined and categorised, which is precisely what is at issue. Until we develop a systematic approach to cognitive ontology, it will be hard to say whether or not we should expect to find one-to-one mappings between functional states and neural structures.

6.1.2 – The cognitive atlas project (Poldrack)

Poldrack (2010) presents a summary of the current state of the art in cognitive ontology research, and argues that current strategies "may be fundamentally unable to identify selective structure–function mappings" (*ibid*: 753), thus calling into question the viability of our current cognitive ontologies. He proposes a more radical revision to our cognitive ontology, based on the utilization of data-mining approaches and a centralized database that he calls the Cognitive Atlas Project. In this section I will present an overview of his approach, along with a summary of his criticisms of previous strategies.

Poldrack opens with a hypothetical comparison of modern neuroimaging with 19th century phrenology. He asks us to imagine that fMRI technology had been developed in the 1860s, and paints a vivid picture of what might have happened:

> Instead of being based on modern cognitive psychology, neuroimaging would instead be based on the faculty psychology of Thomas Reid and Dugald Stewart, which provided the mental ''faculties'' that Gall and the phrenologists attempted to map onto the brain. Researchers would have presumably jumped from phrenology to fMRI and performed experiments manipulating the engagement of particular mental faculties or examining individual differences in the strength of the faculties. They almost certainly would have found brain regions that were reliably engaged when a particular faculty was engaged and potentially would also have found regions in which activity correlated with the strength of each faculty across subjects. (Poldrack 2010: 753)

The point of this somewhat provocative comparison is to draw attention to the reliance of neuroimaging studies on a pre-existing theory of what the relevant functional categories for cognition are. Mappings of some sort between these categories and neuroimaging results will always be possible, but doing so will often involve the kind of problematic one-to-many and many-to-one mappings that Price & Friston reject. Echoing Price & Friston, Poldrack calls for "selective association between mental processes and brain structures" (2010: 754), where each specific structure is associated with a single functional process, and vice versa. He goes on to consider current research strategies, which he argues are unlikely to discover selective associations. There are two main reasons why this might be the case: the underlying ontology might simply be incorrect, or alternatively it might be that the tasks we are using fail to investigate that ontology accurately. (He briefly mentions a third possibility: that functions might map to *networks* rather than discrete structures, but does not discuss it in detail. I will return to this possibility in 6.3.)

Poldrack describes a progression from an early research strategy based around conducting a task and reading off which area it was associated with to a more refined strategy where activation across different tasks is compared in order to determine with greater precision which areas correspond to which functions. He notes that whilst this has led to "increasingly sophisticated functional characterizations of specific anatomical regions" (2010: 756), it has also been the case that single regions have come to be associated with multiple distinct tasks (this is the one-to-many mapping problem raised by Price & Friston). If we continue in this way, he argues, we will end up with a potentially unhelpfully broad definition of which functions each area is associated with, albeit one that is technically valid (see 6.1.4 below for further discussion). The issues here are related to those that I discussed in the previous chapter, as what Poldrack is hoping to discover is essentially the set of genuine cognitive scientific kinds.

In order to develop a more accurate cognitive ontology, Poldrack has set up a central database, which he calls the Cognitive Atlas Project. The aim of this database is to develop "a comprehensive, formally specified ontology of mental processes"

(2010: 756) by comparing correlations between tasks, concepts, and neuroimaging data across a wide range of experiments. Once we have begun to develop this ontology, we can use it to carry out more accurate experiments, and thus further refine our data. The Cognitive Atlas describes connections between conceptual terms such as "working memory", measures used by particular experimental paradigms, and the neural regions that are active when these measures are used. Eventually this would allow researchers to make more accurate predictions about which regions will be associated with which task, and to propose more fine-grained conceptual distinctions to make sense of these predictions. Poldrack describes one early attempt at this kind of process, carried out by Lenartowicz *et al* (2010), which focused on four concepts associated with executive function, and discovered that some of the concepts were more easily dissociated than others. I discuss this study in more detail in 6.2.3.

6.1.3 – After phrenology (Anderson)

Anderson (2014) builds on the work of Poldrack and others to propose an even more extreme revision to our cognitive ontology. He takes Poldrack's phrenology analogy to heart, connecting it to the computational paradigm that has dominated modern cognitive neuroscience, in order to propose a radical departure from this paradigm.

One of Anderson's main concerns with our current ontology is that it fails to properly take into account the phenomenon of neural reuse. There is increasingly good evidence that neural populations are recruited for numerous distinct tasks, even on a relatively small scale (cf. Anderson 2010). This seems to preclude the possibility of identifying discrete one-to-one mappings between functions and regions, leading Anderson to argue that "the prospect of a clear-cut mapping of function to structure appears dim" (2014: 5). According to his analysis, the functional structure of the brain just doesn't lend itself to the kind of one-to-one mapping that previous attempts at ontology revision have been looking for.

This leads him to propose that we characterise neural regions in terms of their *personalities* rather than *function*, where personalities are understood as "the

functional dispositions of individual regions, their underlying causal powers, and their propensities to cooperate with sets of other regions" (Anderson 2014: 114). More technically this proposal involves the generation of multidimensional "fingerprint plots" that represent the full range of functional properties associated with the brain (*ibid*: 118). These fingerprint plots closely resemble the diagrams used to represent human personality traits (see 6.2.2 below), and are intended to predict activation in a region across a wide range of tasks. For example, the plot for the left inferior parietal sulcus shows the most activation on inhibition tasks, somewhat less activation on vision, motor learning, observation, and preparation tasks, and so on. Rather than coming up with a novel functional description that predicts this behaviour, Anderson wants us to give a multidimensional characterisation that accounts for the contributions of this region to a diverse range of tasks.

In contrast with both Poldrack and Price & Friston, Anderson suggests that we should not expect to find one-to-one mappings even after we have revised our ontology. Instead he advocates a move towards a dispositional understanding of neural regions, where we understand the general kinds of processes that a region is involved in, rather than its precise function. One downside to this strategy is that it might seem to lack the kind of sharp explanatory and predictive power that one-to-one mapping could provide, as it is debatable whether dispositions are able to figure in causal explanations (for discussion see Choi & Fara 2016: sec. 6). However, if Anderson is correct it might simply be the best we can do, given the prevalence of neural reuse. We may need to get used to dispositional explanations in cognitive neuroscience, whatever form they might take.

6.1.4 – Context sensitive mappings and multifunctionality (Klein and McCaffrey)

Klein (2012) is presented as a response to Price & Friston, but the arguments that he makes are also relevant to cognitive ontology more generally. His main concern with Price & Friston's approach is that ends up giving extremely vague characterisations of neurocognitive functions. Consider their proposal that we should characterise the left posterior lateral fusiform region as performing sensorimotor integration. Whilst

it is seems clear that this is something that the region does do, it is also what many other regions of the brain do, and this characterisation does not tell us anything about why the left posterior lateral fusiform region *in particular* activates during reading tasks. As Klein puts it, "the most general function that can be attributed to a region is not guaranteed to be cognitively interesting" (2012: 955).

Klein's proposed solution is to make careful use of context specific mappings between functions and regions. He suggests that previous disputes in cognitive ontology and function-region mapping can be attributed to failures to correctly specify the context in which tasks are carried out, along with a more fundamental concern that "we might be profoundly mistaken about which contexts there are" (Klein 2012: 957). If our neuroimaging studies are set up without paying adequate attention to context, it might be unsurprising that we end up with contradictory and confusing results. Klein proposes that we could test our assumptions about the structure of tasks by looking at similarities and differences across related sets of tasks, and cites approvingly Poldrack's (2010) and Lenartowicz *et al*'s (2010) work in this area. He also suggests using the "neural context" of a task, i.e. looking at what other, seemingly unrelated neural areas are activated during a task, as a proxy for task context. So if we find that the motor cortex lights up during what we thought was only a visual processing task, it might indicate that either this task also involves motor processing, or that visual processing of some sort is also carried out in the motor cortex. The resulting picture will not give us a one-to-one mapping between tasks (as proxies for functions) and brain regions, but it might facilitate a deeper understanding of which set of tasks are associated with a particular region. This could in turn lead to a reclassification of our tasks in such a way as to support more precise mappings and predictions.

McCaffrey (2015) raises a similar issue, although he then goes on to argue for a distinct solution. McCaffrey agrees with Klein and Anderson that the apparent multifunctionality of neural regions stands in the way of clear one-to-one mappings between regions and functions. His solution, however, is not to propose a 'one size fits all approach', as Klein does with context sensitivity or Anderson does with

neural personalities, but rather to emphasise what he calls "the brain's heterogeneous functional landscape" (*ibid*). His suggestion is that, just as elsewhere in biology, different kinds of mapping strategy are most appropriate for different cognitive functions and brain areas. He outlines three distinct strategies, along with examples from both general biology and cognitive neuroscience:

Sub-divide and conquer: This strategy is based on Craver's mechanistic decomposition (2001), and is appropriate whenever two functions appear to activate two distinct structures. Here we should distinguish the sub-regions in line with the functions, in order to allow for systematic one-to-one mappings. For example, human pancreatic tissue performs at least two distinct functions: hormone production and digestive enzyme production, but each of these functions is performed by distinct cell populations – so rather than saying that the pancreas has two functions, we should sub-divide the pancreas into discrete cell-populations that each have a single function (McCaffrey 2015: 1016). Similarly, a region of the brain might initially appear to perform several functions, but on closer inspection it could turn out to consist of several functionally distinct structures, individuated in terms of differential activation patterns.

Systematic mapping: This strategy is essentially the same as that suggested by Price & Friston, and is appropriate whenever two or more apparent functions are actually implemented by a single mechanism. Here McCaffrey recommends that we redefine these as a single function, allowing systematic one-to-one mapping between function and mechanism. An example where this strategy is appropriate is a case where a single gene plays the same role in producing two distinct amino acids, or in the case of the brain where the intraparietal sulcus apparently plays the same role (representing analog magnitude) in a number of distinct tasks (McCaffrey 2015: 1017-8). In cases such as these a single higher-level function is "conserved" across tasks, and is thus deserving of a unified place in our ontology.

Context-sensitive: This strategy is comparable to that proposed by Klein, and is appropriate whenever two or more functions are performed by genuinely distinct mechanisms. For example, depending on activity elsewhere in the body, liver hepatocyte cells either absorb glucose or produce bile, two very different functions that cannot be meaningfully united under a single more abstract function (McCaffrey 2015: 1019). In the brain, it seems that the hippocampus is involved in both spatial navigation and episodic memory, and that it performs both of these functions in distinct ways (*ibid*). In both cases it is most useful to say that a single region of the brain or body performs two distinct functions depending on context.

What each of McCaffrey's strategies has in common is an appeal to the underlying mechanistic structure of the functions in question. The sub-divide and conquer strategy appeals to the notion of mechanistic decomposition, and argues that we should treat evidence of distinct functions as evidence of equivalently distinct structures, which can be cashed out as components or sub-mechanisms. The systematic mapping strategy argues in the opposite direction, appealing to the idea that a single mechanism could sometimes be involved in two apparently distinct functions. Finally, the context sensitivity strategy appeals to the idea that when giving mechanistic explanations, we must always have a target phenomenon in mind – in different explanatory contexts, one and the same mechanism might perform two or more distinct functions.

I will return to the topic of mechanistic explanation later in this chapter, where I will suggest that it provides an overarching framework for cognitive ontology formation (McCaffrey himself makes a similar claim). First, though, I will turn to some more detailed case studies of cognitive ontology formation, in order to assess some of the options and strategies that we could adopt in order to revise our cognitive scientific ontology and do away with misleading folk concepts.

6.2 – Case Studies in Cognitive Ontology Formation

In this section I will consider several case studies of cognitive ontology formation, each of which provides useful insights into how we might move forward with cognitive ontology revision. In 6.2.1 I will discuss the development of the popular 'five-factor' model of personality types, which indicates how multidimensional reduction from an initial folk taxonomy can be a way to identify functional scientific kinds. In 6.2.2 I discuss Wierzbicka's attempt to construct a 'natural semantic metalanguage' to help facilitate the identification of cultural universals. Finally, in 6.2.3 I discuss Lenartowicz *et al*'s attempt to refine our understanding of different concepts relating to 'cognitve control'. Whilst only the last study relates directly to the neuroscientific ontologies that are the main focus of this chapter, the first two nonetheless help highlight some general issues surrounding cognitive ontology formation.

6.2.1 – The five-factor model: a case study in cross-cultural convergence

Contemporary personality psychology has settled on a broadly accepted taxonomy of five core personality traits: openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism – the so-called "Big Five". These traits were developed by comparing a wide range of distinct categorisations before settling on what appeared to be the most universally applicable taxonomy, and as such it provides an interesting case study in cognitive ontology formation. The five-factor model attempts to provide a kind of grand-unified theory for personality psychology, with a history dating back to the early 20th century (Digman 1990).

Development of the model began with the work of two German psychologists, Klages (1926) and Baumgarten (1933), who "suggested that a careful analysis of language" (Digman 1990: 418) could help determine accurate personality categories. This work was subsequently systematised by Cattell (1943, 1946, 1947, 1948), who carried out a series of experiments analysing the terms used by college students to describe their peers. Cattell's initial taxonomy was extremely complex, consisting of 30 distinct ratings, but later work in the 40s and 50s settled on five

factors close to those that we have today. Norman (1963) and Smith (1967), amongst others, demonstrated that these categories had impressive predictive power, suggesting that they were tracking something more than mere surface level generalisations.

John & Srivastava (1999) note that the five-factor taxonomy is not committed to any particular theoretical approach, but rather serves to provide a common conceptual landscape for researchers working in personality psychology. This is certainly a pragmatic advantage of the taxonomy, but it does also suggest a potential weakness: does the taxonomy track anything other than behavioural dispositions, and if not, should we treat it as an instrumental tool rather than an accurate description of how the mind works? In the latter case it seems that we would have come no closer to identifying an *accurate* ontology, rather than a merely *useful* one. Later on in this chapter I will discuss the realism/instrumentalism issue in more general terms, as it also applies to other attempts to construct cognitive ontologies.

More recent work on the five-factor model has established that it applies reasonably robustly across cultures and languages (see e.g. Caprara & Perugini 1994, Szirmak & De Raad 1994, Hofstee *et al* 1997), although in each of these studies at least one trait does not seem to apply as well as the others. John & Srivastava conclude that "factors similar to the Big Five have been found in many other languages but often, more than five factors needed to be rotated and sometimes two indigenous factors corresponded to one of the Big Five" (1999: 14). It may be worth noting that most of these studies were conducted in European societies – the evidence from other cultures is somewhat less clear. It is possible that there may be cultural variation in the way that personality traits are expressed, even if the underlying traits or mechanisms are the same. Alternatively, the traits themselves might be culturally specific, in which case any apparent similarities might be due to those features of human existence that are relatively stable across contexts.[41] In any case, the five-factor model has enjoyed considerably more success across cultures

[41] See chapter 2 for further discussion of cultural variation in folk psychology.

than any other attempt at providing a universal taxonomy of personality traits.

The five-factor model has been especially influential in the design of standardized questionnaires, which required a shared taxonomy in order to make comparisons between different questionnaires at all viable. John & Srivastava (1999: 15-17) describe this process in some detail, and note that there remains some disagreement about exactly which set of five traits should be used; although when different labels are used there is still an impressive degree of convergence. They suggest that the traits might be best characterised as prototypes with fuzzy boundaries, rather than absolutely discrete and well-defined categories. Again, whilst this may pragmatically be the best solution, it does call into question the applicability of the model for more precise scientific usage. Ultimately this model is still a work in progress, and we should expect it to be further refined as the field of personality psychology develops.

What this study of the five-factor model demonstrates is that even in a domain as complex and apparently messy as human personality, it can still be possible to agree upon a taxonomy that seems to adequately describe the data. Whether or not this taxonomy actually corresponds to any underlying mechanism is a further question that may simply not be relevant for the purposes of personality psychology, where all that matters is that someone's result on a standardized questionnaire is a relatively good predictor of their performance elsewhere.

The process of refining this taxonomy down to 5 traits from an initial set of 30 is also interesting, and seems like it might generalise to other domains. Similar procedures are applied in neuroscience, for example, where an initial dataset composed of a large number of variables can be reduced down to a more manageable number by applying various dimensionality reduction methods (see e.g. Cunningham & Yu Byron 2014). An interesting feature of these methods is that they can sometimes reveal patterns that would otherwise have remained opaque, allowing for the creation of novel taxonomical categories that go beyond the distinctions suggested by folk psychology. Imagine a case that at first glance seems to feature many-to-many mappings between neural structures and (folk psychologically

defined) functions. By applying some kind of dimension reduction algorithm, we might discover that there is a best fit that allows us to group these disparate functions into new, non-folk psychological categories that map more neatly onto the existing neural structures.[42] In such cases McCaffrey's systematic mapping strategy might be appropriate, but this would not have been apparent prior to applying these methods to the data.

6.2.2 – The Natural Semantic Metalanguage

Wierzbicka's work on comparative linguistics as a guide to cognition was discussed briefly in 2.2.3. Her claim is that "empirical universals of language" can be used as a guide to our understanding of cultural and cognitive universals, and that furthermore "genuine cultural and cognitive universals cannot be formulated" without such a guide (Wierzbicka 2005: 256). To this end she has attempted to construct a "natural semantic metalanguage" (NSM), which "corresponds to the shared lexical and grammatical core of all natural languages" (*ibid*). In terms of folk psychology, if this project were successful then the NSM would contain all of the cultural universal mental state *attributions*. Whether or not these attributions would correspond to actually universal mental *states* is an additional question, although they might at the very least provide a plausible foundation for further investigation.

Each NSM (there is one for each language, corresponding to the core NSM but translated into local concepts) consists of a "mini-language carved out of [the parent language] and based exclusively on empirically established language universals" (Wierzbicka 2005: 258). Whilst the exact expression of these universals will depend on the parent language, they are intended to be equivalent across languages, and are given definitions in basic terms ("semantic primes") that aim to avoid any ambiguity or unwanted implications. For example, Wierzbicka identifies SOMEONE and BODY as two of the 'substantive' semantic primes expressed in

[42] Or, *vice versa*, we could use this strategy to discover new groupings of neural structures that more accurately match our pre-existing functional categories. The choice of which strategy to pursue depends partially on the empirically data, and partially on how revisionary ones attitude towards folk psychology is. I return to this topic towards the end of the chapter.

English, THIS as one of the 'determiner' primes, and HAVE as one of the 'existence and possession' primes (2005: 259). These can then be combined to express one aspect of the (apparently) universal model of a human being: "this someone has a body" (*ibid*: 265).

In this way Wierzbicka has put together what she claims is "a universal folk model of a person", the full details of which can be found in her 2005 paper (2005: 265-6). If this model were an accurate depiction of how people across cultures and languages talk about 'the person', it would at least give us a stable foundation from which to discuss cultural variation in folk psychology, and perhaps even allow some insight into the genuine structure of human cognition and behaviour At the very least, there is something valuable to be said for the way in which Wierzbicka draws our attention towards the bias implicit in constructing theories of language and cognition based solely on English-language predicates and concepts.

However, whilst admirable in scope, the NSM is not ultimately well-suited to the construction of an objective cognitive ontology. Wierzbicka herself admits that according to her account "language doesn't reflect reality directly" (Wierzbicka 1992: 7), although this seems to be in tension with her purported aims. Either the NSM is simply reflective of biological or cultural predispositions towards a certain kind of linguistic system, which is interesting in its own right but not relevant for my purposes, or it picks out genuine features of the world which every language would need to represent (or possibly, but implausibly, it could simply be a coincidence that all languages seem to be composed of these basic elements). It is unclear which of these Wierzbicka has in mind, and this question must be resolved before the NSM can be applied to cognitive ontology revision.

In attempting to reduce language down to a set of primitive meanings, the NSM also opens itself up to the charge of glossing over contextual complexity (Blumcyński 2013: 268). A semantic prime such as BODY is meant to refer to a limited core of what the word might imply in the full context of spoken English, but it is not necessarily so easy to extract meaning from context like this. Wierzbicka contends that the English BODY can be taken as equivalent to the French CORPS

and Polish CIAŁO (Wierzbicka 2010: 17), but it is unclear whether it will ever be possible to rid these words of their original cultural context. Even in an academic context our personal background and cultural context exerts an influence on how we understand what is said, and what inferences (implicit or otherwise) we make, and in idealising away from this complexity NSM is at risk of asserting false (or at least misleading) equivalences.

Nonetheless, the analysis and tools that Wierzbicka presents are valuable, and I will propose something similar to a metalanguage later in this chapter when I discuss a methodology for the formation of an accurate cognitive ontology. This will not be a 'natural' metalanguage, but an 'artificial' one, with the meaning of each word operationalized in terms of experimental measures and concrete, testable claims about the human cognitive system. The end goal, however, is more or less the same as Wierzbicka's: to be able to describe experimental results and theoretical constructions as unambiguously as possible.

6.2.3 – An Ontology for Cognitive Control

Lenartowicz *et al* (2010) present an attempt to apply Poldrack's method for cognitive ontology to the concept of 'cognitive control' (it is worth noting that Poldrack is listed as the corresponding author on this paper). They begin by outlining their general approach, which is to first specify an initial ontology of candidate mental constructs by mining "existing text corpora, such as journal abstracts" (*ibid*: 682). This gives a baseline ontology from which to begin revising with reference to neuroimaging data. In this case they used an earlier study conducted by Saab *et al* (2008), which "isolated a set of five key terms that summarized the literature on executive function" (*ibid*). These terms were "working memory", "response selection", "response inhibition", "task switching" and "cognitive control". Each of the first four terms regularly co-occurs with "cognitive control", and is in turn reliably associated with a number of indicators (behavioural tasks) and heritability measures, suggesting that they already constitute at least a partially accurate ontology. However, the indicators associated with cognitive control are also

associated with the other constructs, suggesting a degree of conceptual overlap. To resolve this ambiguity, Lenartowicz *et al* conducted a meta-analysis of brain activity associated with each indicator task, based on the assumption that genuinely distinct components of the ontology for cognitive control should show distinct activation patterns.

They found that whilst there was a clear distinction between the patterns of activation corresponding to *response selection* on the one hand, and between *working memory*, *response inhibition* and *cognitive control* on the other, there was no clear distinction between the tasks associated with the latter group. The data corresponding to *task switching* was unclear. Based on this data they conclude that response selection is a distinct function associated with the precentral gyrus and middle frontal gyrus, and that cognitive control, response inhibition and working memory may together constitute a second distinct function associated with "a right-lateralized network involving frontal and subcortical regions" (Lenartowicz *et al* 2010: 688). They acknowledge that their data was somewhat noisy, and so do not present these results as conclusive, but do take them to be indicative of the kinds of revision that we should make to our ontology of cognitive control.

One interesting methodological issue that they touch on is the fact that the lack of activation differentiation may be a result of the way in which tasks are associated with functions, rather than a problem with the functions themselves. For example, many tasks targeting task switching may also require response inhibition, even if the experimenters do not explicitly mention the latter. Thus the two functions may appear to overlap because the tasks are not fine-grained enough, rather than the functions themselves being ill formed. In the long run this kind of ambiguity would be resolved by collecting data from a large number of distinct tasks, some of which would hopefully not overlap, but it does prevent any clear conclusions being drawn from Lenartowicz *et al*'s study, even if their overall methodology is a useful one.

6.3 – Methodological Issues and Mechanistic Explanation

In the first half of this chapter I have outlined several basic approaches to cognitive ontology revision, and presented three illustrative case studies. I will now turn to more general methodological issues, which I have previously only mentioned in passing. Resolving these issues will be essential if we are going to be able to settle on a systematic methodology for cognitive ontology revision. I will propose that one way of doing so is to adopt the mechanistic approach to explanation, which allows for non-reductive multilevel integration as a way of accounting for the relationship between different kinds of domain-specific cognitive ontologies.

6.3.1 – Systematic underdetermination

Both Klein (2012) and McCaffrey (2015) raise doubts about the possibility of establishing one-to-one mappings between neural structures and cognitive functions (see section 6.1.4 above). If they are correct then this might seem to call into question the viability of any systematic cognitive ontology revision. The underdetermination originally identified by Price & Friston might turn out to be an intractable feature of our best cognitive neuroscience, rather than a conceptual issue that can be resolved by adopting a novel ontology.

Indeed, there is precedent for this elsewhere in philosophy of science. Stanford (2016) proposes a useful distinction between two different kinds of scientific underdetermination, holistic and contrastive. Holistic underdetermination occurs when a hypothesis cannot be tested in isolation from other hypotheses, whilst contrastive underdetermination occurs when our evidence is equally compatible with two or more distinct theories. Additionally, underdetermination is not an all-or-nothing affair; a theory or hypothesis might be more or less underdetermined. Presumably the global underdetermination posited by Quine would be no worse for cognitive neuroscience than for any other scientific discipline, so if cognitive ontology is interestingly threatened by underdeterminacy it must be in a way that is distinctive to cognitive science.

The underdeterminacy that I am concerned with is contrastive in nature, and relatively limited in scope. Consider the original example given by Price & Friston: the left posterior lateral fusiform region is associated with both visual word processing and animal attribute processing, leaving us with (at least) three distinct hypotheses. The region might have two distinct functions, or it might perform a single function that contributes to both tasks, or it might actually be two distinct regions, one performing visual word processing and the other performing animal attribute processing. The neuroimaging data alone gives us no reason to prefer one hypothesis to the others. However, this underdeterminacy is far from total: we have good reason to prefer a relatively limited range of hypotheses about brain functions, and we can triangulate data from across a range of measures in order to further limit that range. Such triangulation will be a core feature of the methodology that I will describe later in this section.

Whilst Price & Friston conclude that the region is performing sensorimotor integration, this requires additional assumptions about neural organisation, and it is relatively straightforward to make a plausible case for either of the other hypotheses that they mention. Furthermore, as Klein (2012) points out, it is not entirely clear that 'sensorimotor integration' is an explanatorily productive function to attribute to the region. Once we allow mappings at this level of abstraction, it turns out that large areas of the brain can be construed as *performing* sensorimotor integration, and similarly that many tasks can be understood as *requiring* sensorimotor integration. Pushed to its limits, Price & Friston's reasoning might lead us to conclude that all the ever does is sensorimotor integration. Clearly a more detailed story about the function of the left posterior lateral fusiform region is required. McCaffrey points towards one way of constructing such stories, by allowing for a range of distinct mapping strategies, which I will later argue is a core feature of mechanistic explanation.

H. Clark Barrett (2012) suggests an alternative solution to this underdetermination problem, coming from the perspective of evolutionary psychology. He specifically addresses the case of what is traditionally identified as

the visual word form area (*ibid*: 10737), which as noted above appears to be somewhat misnamed as it associated with tasks other than visual word processing. Barrett notes that given the timescale of the evolution of language, it would be strange to find a region devoted to language alone. He proposes that this region is specialised for "category specific object recognition", a function that could be recruited both for visual word processing and other visual processing that require category specific object recognition. Thus a region that predates the evolution of language could plausibly be recruited for visual word processing despite never having been selected for this specific task. This solves the apparent underdetermination problem, i.e. the problem of determining what the visual word form area is specialised for, by proposing a novel (and perhaps more evolutionary plausible) specialisation that accounts for all of the data. His approach also lends itself to the mechanistic account that I will propose later in this section, as he stresses that evolutionary plausible specialisations will be "hierarchically organised", i.e. that complex specialisations can be decomposed into a series of increasingly simple specialisations, perhaps bottoming out in the basic activation profiles of single neurons.

One way to solve the underdetermination problem, I want to suggest, is to acknowledge that we are unlikely to find one-to-one mappings between the kinds of high-level functions that folk psychology is interested in, such as word recognition, and the actual structural organisation of the brain. Rather we should expect to find a hierarchy of increasingly simple functions, each of which can be recruited for multiple distinct tasks and none of which should be expected to easily conform to folk psychological categorisations. This cuts both ways, in the sense that a single (folk psychological) task, such as word recognition, is plausibly going to recruit functions associated with numerous neural structures, but also that a single neural structure, such as the left fusiform gyrus, is plausibly going to perform a function, such as category specific object recognition, that is recruited for several distinct tasks. Approached in this way, underdetermination because a feature of the

relationship between tasks defined at the personal level and functions specified at the neural level, not a negative consequence of our lack of understanding.

6.3.2 – One ontology or many?

Each of the approaches that I discussed in 6.1 would lead to slightly different revisions to our ontology. If we followed Price & Friston's original suggestion we might end up with a small number of very general functions, such as 'sensorimotor integration'. Poldrack's Cognitive Atlas project aims to uncover a large number of finely individuated states, processes, and functions, although it could take a long time and a large quantity of data before these settled into any kind of stable ontology, if indeed there is a stable ontology to be found. Alternatively, we might turn to Anderson's reclassification of neural functions in terms of state spaces, which would consist of a core ontology of 'personality types' with which each distinct region could be classified. Klein and McCaffrey both advocate choosing different strategies for ontology revision depending on the context (in Klein's case) or on the region/function in question (in McCaffrey's case). Is one of these approaches clearly superior? If not, can we reconcile the different approaches, or should we accept a pluralistic approach to cognitive ontology revision?

As suggested in the previous section, what I want to propose is that there is no incompatibility between tasks defined in folk psychological terms at the personal level and functions defined according to some other criteria at the neural level. Relatedly, the criteria we choose for individuating neural level functions might depend somewhat on whatever it is that we are currently interested in explaining. This is not to say that there is no determinate answer to the question of what function a region performs, but rather that there might be a hierarchy of regions and functions that allows for some selective interpretation of which particular function is most relevant at any given time.

The analogy that I made with computational individuation in section 4.2.3 will once again be useful here. Shagrir (2001) presents a case where a computational system is sensitive to three different categories of voltage level, and yet can be

interpreted as performing the logical function AND, which is sensitive to only two kinds of logical category (TRUE and FALSE). Should we say that the 'ontology' of this system consists of three states (according to voltage level) or two states (according to logical function)? In response to this puzzle I have argued that whilst ultimately the mechanistic structure of this system forces us to recognise three distinct states, it can easily be interpreted as instantiating two states, and there is no incompatibility between acknowledging the pragmatic value of this interpretation whilst also recognising the existence of the underlying three-state mechanistic structure (see Dewhurst 2016). Analogously, whilst I think it is important that we should eventually agree upon a single basic ontology that reflects the actual mechanistic structure of the brain, this is not incompatible with the existence of higher-level ontologies that are posited relative to our explanatory interests. This is a pluralism of sorts, but if Craver (2012) is correct, then it is an explanatory pluralism that is inherent to the project of mechanistic explanation, and not one that in any way calls into question the 'objectivity' of our scientific explanations. I will expand on this further in the next subsection, where I outline a mechanistic approach to the cognitive ontology project.

6.3.3 – Mechanistic explanation and multilevel integration

Over the past two decades mechanistic explanation has come to be seen by many as the primary mode of explanation in cognitive science (see Craver & Tabery 2016 for an overview). In this section I will argue that adopting the mechanistic account of explanation is the best way to make sense of cognitive ontology revision, as it allows for a principled non-reductive stance towards the different levels of our ontology. Furthermore, adopting the mechanistic model of explanation will allow us to clarify the relationship between folk psychological discourse and cognitive scientific explanations. I will argue that folk psychology can sometimes provide sketches of mechanisms, which can be a useful starting point for scientific investigation, but that will eventually have to be replaced with full mechanistic explanations. This will

further support my claim that we can revise our cognitive ontology without needing to eliminate folk psychological discourse.

Mechanistic explanation is typically contrasted with the traditional deductive-nomological model of explanation, which attempts to discover 'covering laws' that will apply universally across a given domain. For example, given the laws of Newtonian physics and some data about the current positions, masses, and velocities of the objects in the solar system, we can deduce to a fairly high degree of certainty what the future positions of these objects will be (Woodward 2014: 2.1). Explanation of this sort works well in the physical sciences, where generalisable laws are relatively easy to discover, but it is less well suited to the special sciences, where exceptions and special cases abound. Whilst in theory it would be possible to posit a large enough number of well-defined general laws to cover every case, in practice this has proved unfruitful. In contrast, mechanistic explanation proceeds by positing a particular mechanism whose components interact to produce a given phenomenon, rather than attempting to cover every case all at once.

A mechanism is defined by Glennan as 'a complex system that produces the target phenomenon by the interaction of a number of parts' (paraphrased from Glennan 1996 and 2002). The 'target phenomenon' is simply whatever we, or the community of scientists that we are studying, are interested in explaining. Typically an initial functional analysis provides a basic sketch of the system that produces the target phenomenon, which is subsequently refined through experimentation until a full mechanistic explanation can be given (Piccinini & Craver 2011). The system (i.e. the mechanism) will be broken down into interacting components, thus revealing its causal structure. It is the interaction of these components that produces the target phenomenon, and taken as a whole this process provides an explanation of how the target phenomenon is produced.

Mechanistic explanation, strictly speaking, is just a special kind of causal explanation. The contribution that each component makes towards the production of the target phenomenon is typically cashed out in causal terms, and ultimately it should be possible to fully decompose the mechanism into basic relations that can be

explained in terms of chemical or physical laws. However, this does not mean that mechanistic explanation is reductive. The production of the target phenomenon should be understood as a function of the mechanism as a whole, and mechanistic explanation can span multiple levels, as discussed by Piccinini & Craver (2011).

In the case of cognitive ontology revision, this framework gives us the tools necessary to explain the relationship between different levels of the ontology. To return again to Price & Friston's original example, there might be one high-level sense in which the left posterior lateral fusiform region is best described as performing sensorimotor integration, but another more specific sense in which it performs word recognition. Craver (2012) has argued that mechanistic functions should be individuated in terms of an explanatory perspective, which would allow for distinct functional ontologies for different branches of cognitive neuroscience. One implication of Craver's account is that there is no single, determinate function performed by each mechanism, but rather a multitude of context-specific functions. This is no bad thing. As Hochstein (2016b) demonstrates, a single mechanism can contribute to several distinct explanatory contexts. For instance, this allows us to account for the fact that the left posterior lateral fusiform region can be attributed different functions depending on the task (i.e. the target phenomenon) being investigated.

More generally, the way in which neural mechanisms are individuated will have to be sensitive to whatever it is that we are currently trying to explain. This does not meant that we won't be able to give a single unified description of the neural system as a whole, but rather that until we know what we are trying to explain, there is no objective sense in which we can say what the function of any part of this system is. Think of it like this: what we have before us is an extremely complex mechanism, and even if we can explain in great detail what the structure of this mechanism is and how it all fits together, we still won't have answered the additional question of what it is for. Once we understand the structure though, we can put the mechanism in a situation (i.e. ask it to perform a task), and see how it responds to that situation. At this point we can begin attributing functions, and begin

saying, for instance, that when this mechanism is engaged in a reading task then such-and-such a component of the mechanism performs a word recognition function. In another context, say when the mechanism is performing a hunting task, the very same component might perform an animal recognition function. Barrett's evolutionary psychology project (discussed in 6.3.1) could also be construed as an explanatory context in its own right, from the perspective of which this component might be best understood as a 'category specific object recogniser'.

6.4 – The Contribution of Folk Psychology to Cognitive Ontology Revision

Adopting the mechanistic explanation framework described in the previous section allows us to reconcile the revision of our folk ontology for scientific purposes with our retention of it for everyday usage. In order to do this I will argue that folk psychological descriptions can sometimes function as sketches of mechanisms, giving a broad-brush picture of the phenomenon that we are trying to explain. This means that applying folk psychological concepts to cognitive scientific explanations constitutes a sort of category error – it would be comparable to using a common-sense description of a ball falling to the ground when dropped in order to explain *why* the ball falls to the ground when dropped. If we were to try and explain this phenomenon by appealing to the unanalysed folk notion of 'dropping', we would rightly be accused of offering a circular explanation. Here it is important to distinguish a description of a phenomenon from a genuine explanation, where an explanation must in some sense go beyond the initial description in order to be genuinely explanatory. Folk psychology can sometimes give genuine explanations, but these are posited at the level of persons rather than mechanisms, and constitute a distinct mode of explanation to that given by cognitive science. When it comes to genuinely cognitive scientific explanations, we should aim to treat folk psychological description as useful sketches of the phenomenon to be explained, which can give suggestive hints at the form our explanation should take, but nothing more than that. Of course, it will not always be the case that the folk psychological description can provide any useful guidance, for example if it identifies a purely social phenomenon

with no corresponding cognitive scientific mechanism. However, given that we may not be able to identify these cases ahead of time, it will often be worthwhile attempting to treat the folk psychological description as a mechanism sketch, even if ultimately this turns out not to be a useful exercise. The key point is that the folk psychological description of a phenomenon is only one source of data, and should be cross-referenced with other sources in order to uncover the actual structure of the underlying mechanisms.

In the second half of this section I will describe how the process of cognitive ontology revision can draw on data from across multiple disciplines, including those that study folk psychological concepts and intuitions. I will also present a preliminary account of how radical our cognitive ontology revisions might be, although ultimately whether or not this account turns out to be accurate will be an empirical matter. I present it primarily as an exercise in exploring the potential limits of cognitive ontology revision, rather than a definitive proposal, as the empirical evidence is currently too limited to make any conclusive predictions.

6.4.1 – Folk psychological descriptions as sketches of mechanisms

Piccinini & Craver (2011) describe how functional analyses of cognitive systems can be interpreted as giving what they call "mechanism sketches" – i.e., incomplete outlines or "elliptical descriptions" of full mechanistic explanations. By functional analysis they have in mind a classical approach to psychological explanation that focuses on the functional properties of a system, and which maintains that those properties are distinct from their physical implementation (*ibid*: 286ff). Whilst this classical approach will allow that every functional analysis must ultimately be implemented by a physical mechanism, it denies that functional explanation is reliant on mechanistic explanation, or even constrained in any way by the details of the implementation. Piccinini & Craver argue that functional analyses and mechanistic explanations do in fact constrain one another in interesting ways, and that the autonomy and distinctness claims are therefore false. According to them, there is really only one kind of (scientific) psychological explanation – mechanistic

explanation – and functional analyses are only explanatory insofar as they contribute to full mechanistic explanations.

In line with this approach, I think that we should treat (some) folk psychological descriptions as high-level mechanism sketches, in the sense that they can usefully be interpreted as providing functional analyses of cognitive systems. It is important to note that this treatment will not apply to all folk psychological descriptions, but only to those that appear to be attempting to describe sub-personal processes that properly fall into the domain of scientific psychology. In the next section I will outline a way of distinguishing personal from sub-personal explanations, and argue that folk psychological descriptions that fall into the latter domain should be interpreted as mechanism sketches.

One advantage of this treatment of folk psychological descriptions is that it allows us to acknowledge that such descriptions are often incomplete or inaccurate, without having to commit to a fully reductionist or eliminativist position. As Piccinini & Craver are keen to stress, mechanistic explanation is not reductive, and interpreting a functional analysis as a mechanism sketch does not imply that that analysis is false – it is simply incomplete. Furthermore, functional analyses can serve a useful explanatory role by constraining the range of possible mechanistic implementations. So interpreting folk psychological descriptions as mechanism sketches allows us to accommodate them into our explanatory practices as a useful guiding heuristic.

Consider an illustrative example. A folk description of what happens when I duck to avoid a flying object might go something like this:

> He saw the duck flying towards him, and didn't want to get hit, so he ducked to avoid it.

In one sense this description is perfectly accurate, and even seems to provide an adequate explanation of what happened. I did see the duck, and I didn't want to get hit, so I ducked to avoid it. In the next sub-section I will characterise this mode of

explanation as 'personal level'. However there is another sense in which this description doesn't really explain anything at all: all it has done is re-describe what happened in psychological terms, without providing any causal mechanism that might account for my ducking. Nonetheless, if we treat this description as a mechanism sketch it might provide a useful starting point for developing a full (sub-personal) explanation:

> He saw the duck flying towards him [visual system], and didn't want to get hit [motivational system], so he ducked to avoid it [motor system].[43]

We have now begun the process of mechanistic decomposition, using the initial sketch provided by the folk description to determine that there might be three distinct sub-systems involved. From here we can draw on other theories of cognitive processing to posit further sub-components and sub-sub-components (etc.) that might make up the systems in question, and eventually we will have a full mechanistic explanation of why I ducked to avoid the duck. Of course, the folk description might turn out to be unhelpful, if for example the distinction between motivational and motor systems turns out to be misleading. This will become apparent during the process of decomposition and testing, and at the very least the sketch provided by the folk description will have given us somewhere to start our investigation. The possibility of error, in both folk and scientific explanations, is why it is important that our methodology allows for convergence across multiple disciplines, as I will describe in section 6.4.3.

6.4.2 – Personal level explanations and sub-personal sketches

The personal/sub-personal distinction was originally introduced by Dennett (1969), "as a distinction between two ways of explaining human behaviour" (Drayson 2014, see 2.1.1 for further discussion). Personal level explanations are those that are

[43] It is important to note that this is a deliberately scientistic description that is not intended to capture every nuance of the situation.

familiar to us from everyday, folk psychological accounts of behaviour, whilst (for Dennett) sub-personal explanations are those that attribute psychological or mentalistic capacities to parts of people, rather than people as a whole. Whilst I do not want to endorse the straightforward attribution of folk psychological states and processes to parts of people (such as brain states), I do think that Dennett identifies a useful distinction between two quite different modes of explanation.

For my purposes, personal level explanations are those given by folk psychology, whilst sub-personal level explanations are those given by cognitive science. These two modes of explanation are distinguished primarily by what counts as a genuine cause. For sub-personal explanation, the only causes are those that can be associated with physical mechanisms, whilst personal level explanations also allow rational or normative causes. A rational or normative cause is one that appeals to a person's rationality or norm-compliance as a reason for some event occurring. So to return to the example used above, to say that I ducked because I didn't want to get hit is an appeal to a rational cause, which might serve as a perfectly adequate personal level explanation, but can only provide an initial sketch of a sub-personal mechanism. This initial sketch could, for example, posit that there is some kind of mental state corresponding to my desire not to be hit, without properly spelling out what that state is like, or how it is able to cause behaviour. The task of cognitive science, given this initial personal level sketch, is to either uncover the sub-personal mechanisms corresponding to 'desire not to be hit', or else to posit some other mechanism that produces the personal level behaviour, if it turns out that the initial sketch was inaccurate.

Personal level explanations can also invoke physical causes, such as if someone was to say that I fell over because I was hit. Whilst it invokes a physical cause, this explanation still only provides a sketch of a mechanism, as it has not yet spelled out in detail what happens when I am hit, and why that causes me to fall over. In this case, however, the full mechanistic explanation that we might eventually reach would probably be better classed as belonging to the domains of physics or

anatomy, rather than that of cognitive science, as the explanation would have been much the same if I were just an inert lump of flesh (i.e. not a cognitive system).

The upshot of this distinction is that whilst for the purposes of scientific psychology we must treat folk psychological descriptions as mere sketches of full mechanistic explanations, we can nonetheless admit that for the distinctive personal level domain, folk psychology does sometimes give full explanations. The explanations given by folk psychology just constitute a distinct mode of explanation to those given by scientific psychology. This is not a particularly novel observation, but it does help explicate the error that I think is made when folk psychological concepts are applied to cognitive scientific explanations: these concepts are typically appropriate for the personal rather than the sub-personal mode of explanation (although see Figdor 2014 for a contrasting opinion).

6.4.3 – Convergence across disciplines

As we saw in the case of the OCEAN model and Wierzbicka's NSM, it is possible to use a broad survey of folk intuitions as a way to get an initial sense of the layout of a target domain. Turner (2012) also advocates a strategy of this kind as part of a combined, interdisciplinary effort to revise our cognitive scientific taxonomy. Coupled with the data mining strategies outlined by Poldrack (in cognitive neuroscience) and proposed by Apicella & Barrett (2016) for the domain of psychological anthropology, we can begin to see what a unified strategy for interdisciplinary convergence would look like in practice.

Figure 6.2 (below) gives a rough idea of the convergence strategy that I have in mind. Folk psychological descriptions can provide an initial sketch of the target domain, whilst public databases collating data from both cognitive neuroscience (Poldrack's Cognitive Atlas project) and cross-cultural psychology (Apicella & Barrett's as-yet unrealised proposal) will allow for cross-referencing. Theoretical analysis is still required in order to bring all of this together, but it should be constrained by the experimental evidence. As our conceptual taxonomy is updated

and revised we can create better experimental designs that allow for more accurate data, leading to further revisions, and so on, in an on-going, iterative process.

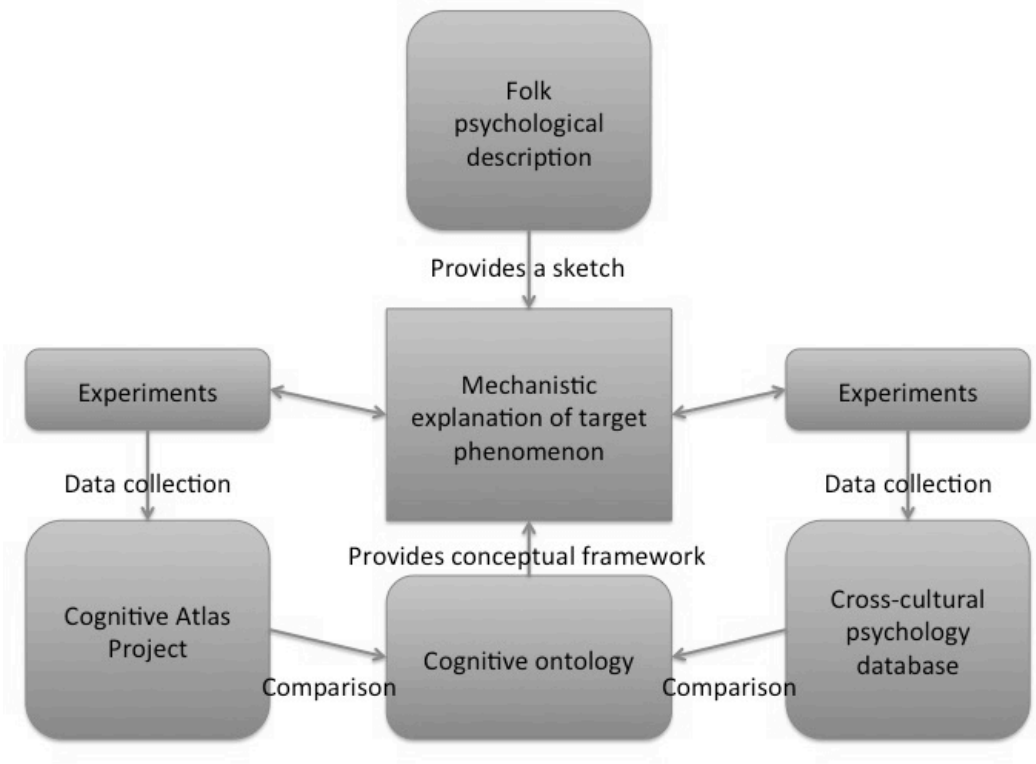

Figure 6.2. A convergent strategy for cognitive ontology revision.

This kind of convergent strategy is important if we want to avoid parochialism in individual disciplines. Without comparing your results with those from other cognitive scientific disciplines, you can risk reifying a specialised ontology that works perfectly for your own theoretical purposes but is incompatible with those of other theories (cf. Sullivan 2016). However, this does not mean that the conceptual distinctions that apply in one context will necessarily apply in others – a useful analogy here is the situation in biology, where distinct taxonomical classifications of species can exist alongside one another whilst still acknowledging that there are facts about the world that remain true across all taxonomies (Ludwig 2015: 48-55). The balance between achieving a level of objectivity and respecting

domain-specific ontologies is a difficult one to achieve, but adopting some sort of unified approach to ontology formation will make it somewhat easier.

Convergence need not be incompatible with a certain kind of pluralism either. Earlier in this chapter I suggested that we should accept that the correct mapping between function and structure will depend to some extent on explanatory context. This is compatible with there being some general consensus with regard to the kinds of functions and structures that there are, even if they are grouped differently by different disciplines. Partly this is a question of grain: for some purposes we might be interested in mapping entire cortical structures, whilst in others we might want map specific voxels, or even individual neurons. So, for example, we might all come to accept a version of Anderson's neural personality-based ontology, but nonetheless carve up this ontology differently depending on the grain of the phenomenon that we are currently investigating. The same goes for functions: in one context 'reading' will be a sufficiently detailed description of the task being investigated, whilst in another we might need to specify that it is word or letter recognition that we are most interested in.

6.4.4 – How radical?

The only question that remains is what our revised cognitive ontology is actually going look like. Whilst this is ultimately an empirical question, depending on both experimental results and to some extent sociological conditions, I will hazard some rough predictions as to how it might turn out.

If Anderson is correct, then we ought to move away from an ontology populated by discrete functional states and processes, and towards one populated by regions that have multiple functions depending on how and when they are recruited. The labels attached to these region would have to move beyond those provided by folk psychology – rather than having a region devoted to memory or belief formation, we might have a region that is associated with a range of tasks including memory and belief formation, but that is also associated with other tasks, perhaps semantic visual processing and intentional motor action. Whilst the tasks described

above can plausibly be individuated in folk terms, this will not be the case for the region, which we would be better off inventing an entirely new term for, either thematically (visuo-motor-memory region), or anatomically (voxel 174). The latter seems like it might be less confusing, although we see both kinds of naming convention at work in contemporary cognitive neuroscience.

However, if we were to adopt something more like Price & Friston and Poldrack's proposals, then the revisions to our ontology might be somewhat less radical, at least in the sense that we would still be looking for on-to-one mappings between structures and functions. Nonetheless, the functions in question would be increasingly divergent from the folk ontology, as they would follow the functional architecture of the brain rather than that of observable behaviours. Price & Friston's proposal for the localisation of the function 'sensorimotor integration' provides one example of this, as does the case study presented by Lenartowicz *et al*.

Finally, we could adopt a version of the pluralism advocated by McCaffrey. In this case the revisions to our ontology might be even less radical still, as we could concede that in some contexts folk psychology concepts turn out to capture useful distinctions, whilst in others they do not. Following my discussion of folk kinds and natural kinds in the previous chapter, this is ultimately the strategy that I believe will prove most successful. In psychology and cognitive science, much like in biology, we should not expect to find a single, discrete set of taxonomical principles that apply across all domains. Nonetheless, it is important to acknowledge that the folk ontology is often incapable of capturing the fine-grained distinctions required for contemporary cognitive scientific enquiry, especially in domains such as cognitive neuroscience and psychophysics.

6.5 – The Relationship Between Folk Psychology and Cognitive Science

In this chapter I have reviewed several recent proposals as to how we should go about revising our cognitive ontology (6.1), focusing on discussions of how best to interpret the relationship between psychological tasks and neurological functions. I then briefly considered three historical cases of cognitive ontology revision (6.2),

each of which provides useful lessons. In section 6.3 I introduced and addressed several related methodological issues, before proposing that these issues can be at least partially resolved by adopting a mechanistic framework for cognitive ontology revision. Finally I considered how folk psychology might be able to contribute to such revision (6.4), concluding that folk psychological descriptions can serve as initial sketches of mechanisms, and can also help to enable convergence across disciplines. I ended that section by offering some suggestive comments as to how radical our eventual cognitive ontology revision might be.

Before moving on I want to say a little more about how I envision the relationship between folk psychology and cognitive science, both in order to clarify what I have said previously and also to assuage some concerns that might arise. As I have indicated previously, the kind of elimination or revision of folk psychological concepts that I am arguing for here is strictly limited to the cognitive scientific domain, and says nothing at all about the use of such concepts in the folk psychological domain. However, it is not always clear where to draw the line between these two domains. Construed broadly, cognitive science includes the likes of anthropology and social psychology, where folk psychological phenomena are sometimes exactly what is being studied. When studying personal level interactions that involve the attributions of mental states, for instance, surely folk psychological concepts are suitable or even necessary? In fact there is no great puzzle here; my concern is only with the *inappropriate* usage of folk psychological concepts in cognitive science, and so in a case like this, where their usage is demonstrably appropriate, there is no issue. Similarly, if it turned out that there was a low level neural state with a functional profile very much like that of belief, I would not necessarily be opposed to the term 'belief' being used.

In the previous section I suggested that folk psychology could sometimes be seen as providing a sketch of a mechanism, in the sense that it identifies a target phenomenon and gives some rough suggestions as to how that phenomenon might be produced. Whilst this might seem to concede that folk psychology can offer some insights into the structure of the underlying mechanism, it is also important to

recognise that a mechanism sketch is often crucially incomplete. Consider an analogy with folk physics: an intuitive sketch of what happens when you release a weight that is being spun on the end of a piece of string is roughly correct about the fact that the weight will fly off in some direction, but (typically) completely wrong about which direction that will be (cf. Churchland 1979; Clark 1987 also suggests a similar analogy). Analogously, whilst a folk psychological sketch of some cognitive process is likely to be roughly correct about the coarse grained behavioural outcomes, it is unlikely to say anything at all about the sub-personal processes that generate those outcomes. It is also likely to get some of the more fine-grained details wrong: if I ask you to predict what I will do if I see that it is raining outside, you might say that I will pick up an umbrella, but actually I am more likely to put on a coat. For most day-to-day purposes these fine-grained details do not matter in the slightest (i.e. you would be able to predict that I will take *some* preventative measure to avoid getting rained on), but it is important to acknowledge that folk psychology does not always give us an entirely accurate picture of human behaviour – whilst also admitting that scientific psychology is itself often no better in this regard. What scientific psychology is good at is giving explanations at the sub-personal level, whilst what folk psychology is good at is giving coarse-grained behavioural predictions in everyday situations (and even these may be somewhat dependent on the mindshaping mechanisms described in chapter 3). This means that whilst folk psychology is good at predicting, it is less good at explaining, at least in the mechanistic sense that I have been describing in this chapter (folk psychology may well be good at giving normative or narrative explanations, but I take it that these are not typically the kinds of explanation that scientific enquiry aims at).

To sum up, folk psychology can serve as an initial guide to the kinds of phenomena that cognitive science is interested in studying, and in this sense can sometimes provide preliminary mechanism sketches. However, it is important to take the conceptual ontology provided by folk psychology with a pinch of salt, as we have good reason to think that the way this ontology carves up the world is not appropriate for many areas of cognitive science. This is because the folk psychological ontology

is heavily influenced by cultural and linguistic factors, and reflects socially constructed kinds that, whilst in a sense just as real as the kinds of a mature cognitive ontology, can only appropriately be applied within the social domain. To this end we should adopt some version of the methodology that I described in this chapter, revising our initial folk psychologically derived ontology with reference to data from across the cognitive sciences, with the aim of developing a complex, multi-level ontology that more accurately captures the mechanistic structure of cognitive systems.

Conclusion: From Folk Psychology to Cognitive Ontology

In this thesis I have aimed to elucidate the complex relationship between two different kinds of discourse about human behaviour, folk psychological on the one hand and cognitive scientific on the other. I began by clarifying the many different ways in which the term folk psychology has been used in the past, and argued that we should conceive of folk psychological discourse as a complex social practice, including not only mental state attributions, but also behavioural predictions, narrative explanations, and regulative constraints. Contrary to the received view in philosophy, this discourse varies across cultures, and thus equating folk psychology with the attribution of beliefs and desires gives a very limited perspective on the phenomenon. I then argued that folk psychology is not primarily in the business of explaining the sub-personal or mechanistic structure of human cognition, and that we can explain its predictive and explanatory success without being committed to it having any scientific basis or validity. By the end of the first half of my thesis I hoped to have articulated a strong positive account of folk psychology as an everyday, commonsense discourse that stands by itself and does not require (or ask for) any vindication from cognitive science. This positive account was intended to undermine the classical debate between realists (such as Fodor) and eliminativists (such as Churchland), and to therefore allow for a more nuanced discussion of the role played by folk psychological concepts in philosophy and cognitive science. Wilkes (1991), Botterill & Carruthers (1999), and others have all suggested a similar response to debates about the relationship between folk psychology and scientific psychology, but I hope that I have been able to add to this by elucidating in more detail the positive role played by folk psychological discourse.

The second half of this thesis began with an initial look at how folk psychological concepts might struggle to capture the full complexity of several contemporary debates in philosophy and cognitive science, including the false belief task in social cognition, the individuation of the senses, the extended cognition debate, and the emerging predictive processing paradigm. In each case I advocated a

disambiguation strategy where coarse-grained folk psychological concepts are replaced with novel, finer-grained concepts that better capture the complexity of the domain being studied. I then proceeded to look in more detail at the status of natural kinds terms in psychology and cognitive science, and argued that whilst folk psychological terms constitute 'human kinds' that are suitable for describing whole persons, they cannot be expected to fulfil the role of natural kinds in a mature cognitive science. Finally, I looked at recent work on ontology revision in cognitive science, and proposed an integrated strategy for developing novel cognitive ontologies. This strategy respects the role played by folk psychology in identifying target phenomena and providing inspiration, but seeks to eventually replace folk psychological concepts with a more empirically motivated taxonomy drawing on data from across the cognitive sciences. In closing I suggested a few possible directions that this taxonomy might take, although ultimately I think this is an empirical matter, the results of which cannot be determined *a priori*.

Whilst writing this thesis I discovered that Jenson (2016) appears to have independently reached a conclusion very much like my own, albeit focusing on the more restricted case of belief. He advocates a 'scientific eliminativism', arguing that "belief is not an appropriate category for cognitive science" (*ibid*: 967), whilst rejecting the 'old-school eliminativist' project of eliminating belief (or folk psychology) from everyday use. His argument focuses on robustness and fragility – theoretical entities, he argues should, should be robust, meaning that their existence can be confirmed by a number of distinct methodologies or measurements. "A theoretical entity is fragile", on the other hand, "if the results of multiple, independent, putatively reliable measures of that entity turn out to radically vary and this variation cannot be adequately explained away" (*ibid*: 969). Belief, according to Jenson, is a fragile theoretical entity, and should thus be eliminated from our scientific ontology. Our agreement on the status of belief (and perhaps other folk psychological concepts) in cognitive science could itself be taken as a sign of the robustness of what Jenson calls scientific eliminativism. Insofar as we have both reached this conclusion via distinct methodologies, it seems plausible that we are

beginning to home in on a genuinely serious concern with the use of folk concepts in cognitive science.

Whilst it may sometimes seem as though I am proposing an extremely reductionist or scientisitic approach towards the study of the mind, this is not at all my intention. Rather I think it is important to recognise the value of two distinct ways of looking at the world, neither of which is more correct or objective than the other. We truly are persons with beliefs and desires, it just turns out that those beliefs and desires are more akin to character traits, like being brave, than they are to discrete functional units, like 1s and 0s in a digital computer. At the same time, we are also complex physical systems composed of interacting parts whose functions often defy folk description. This should not dishearten us – in fact I think this way of conceiving of folk psychology and cognitive science is far more humanistic and intuitive than the realist reading of propositional attitude psychology, where we either end up identifying beliefs and desires with functional states of a computational architecture, or else eliminating them from our ontology once it turns out that we cannot identify any such states. Reconceiving of folk psychology and cognitive science as discourses with distinct domains of enquiry allows us to preserve the autonomy of our everyday interactions and understanding of human behaviour, whilst continuing to uncover the sub-personal mechanisms underlying that behaviour. A cloud does not become any less real once we learn it is composed of water vapour, and a person does not become any less a person once we discover the mechanisms that are responsible for their behaviour.

References

- Adams, F. & Aizawa, K. 2001. "The bounds of cognition." *Philosophical Psychology* 14: 43-64.
- Adams, F. & Aizawa, K. 2005. "Defending non-derived content." *Philosophical Psychology* 18: 661-9.
- Adams, F. & Aizawa, K. 2010. "Defending the Bounds of Cognition." In Menary (ed.), *The Extended Mind*. Cambridge, MA: MIT Press.
- Ahluwalia, A. 1978. "An intra-cultural investigation of susceptibility to "perspective" and "non-perspective" spatial illusions." *British Journal of Psychology*, 69: 233-241.
- Ahn, W., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., Coley, J. D., & Shafto, P. 2001. "Why essences are essential in the psychology of concepts." *Cognition* 82: 59-69.
- Amedi, A., Jacobson, G., Hendler, T., Malach, R., & Zohary, E. 2002. "Convergence of visual and tactile shape processing in the human lateral occipital complex." *Cerebral Cortex,* 12: 1202–1212.
- Anderson, M. 2010. "Neural reuse: A fundamental organizational principle of the brain." *Behavioral and Brain Sciences* 33/4: 254-61.
- Anderson, M. 2014. *After Phrenology.* Cambridge, MA: MIT Press.
- Anderson, M. 2015. "Mining the Brain for a New Taxonomy of the Mind." *Philosophy Compass* 10/1: 68-77.
- Andrews, K. 2008. "It's in your nature: A pluralistic folk psychology." *Synthese*, 165/1:13–29.
- Andrews, K. 2015. "The Folk Psychological Spiral: Explanation, Regulation, and Language." *The Southern Journal of Philosophy,* 53: 50-67.
- Andow, J. 2015. "How 'Intuition' Exploded." *Metaphilosophy*, 46/2: 189-212.
- Apicella, C. L. & Barrett, H. C. 2016. "Cross-cultural evolutionary psychology." *Opinion in Psychology* 7: 92-97.

- Apperly, I. & Butterfill, S. 2009. "Do Humans Have Two Systems to Track Beliefs and Belief-Like States?" *Psychological Review* 116/4: 953-70.
- Astuti, R. 2014. "Implicit and Explicit Theory of Mind." *Anthropology of This Century*. Published online: http://aotcpress.com/articles/implicit-explicit-theory-mind/
- Atran, S. 1998. "Folk Biology and the Anthropology of Science: Cognitive Universals and Cultural Particulars." *Behavioral and Brain Sciences* 21: 547–69.
- Baillargeon, R. 2004. "Infant's reasoning about hidden objects: evidence for event-general and event-specific expectations." *Developmental Science,* 7/4: 391-414.
- Baker, L. 1999a. "Folk Psychology." In Wilson & Keil (eds.), *MIT Encyclopedia of Cognitive Science.* Cambridge, MA: MIT Press.
- Baker, L. 1999b. "What is this thing called 'commonsense psychology'?" *Philosophical Explorations* 2/1: 3-19.
- Barkow, J., Cosmides, L. & Tooby, J. 1992. *The Adapted Mind.* New York: OUP.
- Barrett, H. C. 2012. "A hierarchical model of the evolution of human brain specializations." *Proceedings of the National Academy of Sciences,* 109: 10733-10740.
- Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., Bolz, M., Henrich, J., Setoh, P., Wang, J., Laurence, S. 2013. "Early false-belief understanding in traditional non-Western societies." *Proceedings of the Royal Society B.* DOI: 10.1098/rspb.2012.2654
- Barrett, H.C., Bolyanatz, A., Crittenden, A.N., Fessler, D.M.T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B.A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. 2016. "Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment." *Proceedings of the National Academy of Sciences*, 113/17: 4688–

4693.

- Barrett, L. F. 2006. “Are Emotions Natural Kinds?” *Perspectives on Psychological Science* 1/1: 28-58.
- Barrett, L. F. 2012. “Emotions Are Real.” *Emotion* 12/3: 413-29.
- Baron-Cohen, S., Leslie, A., & Frith, U. 1985 “Does the autistic child have a ‘theory of mind’?” *Cognition* 21: 37-46.
- Baumgarten, F. 1933. “Die Charaktereigenshaften.” In *Beitrage zur Charakter- und Per sonlichkeitsforschun*. Bern: A. Francke
- Beer, R. 1995. “A dynamical systems perspective on agent-environment interaction*.” Artificial Intelligence,* 72: 173-215.
- Bennett, J. 1978. “Some Remarks About Concepts.” *Behavioral and Brain Sciences,* 1: 557-60.
- Bickle, J. 2016. “Multiple Realizability”. In Zalta (ed.), *The Stanford Encyclopedia of Philosophy.*
- Block, N. 1978. “Troubles with functionalism.” In Savage (ed.), *Perception and Cognition.* Minneapolis: University of Minnesota Press.
- Blumczyński, P. 2013. “Turning the tide: A critique of Natural Semantic Metalanguage from a translation studies perspective.” *Translation Studies,* 6/3: 261-76.
- Botterill, G. 1994. “Belief, Functionally Discrete States, and Connectionist Networks.” *The British Journal for the Philosophy of Science* 45/3: 899-906.
- Botterill, G. 1996. “Folk psychology and theoretical status.” In Carruthers & Smith (eds.), *Theories of Theories of Mind.* Cambridge, UK: CUP.
- Botterill, G., & Carruthers, P. 1999. *The Philosophy of Psychology.* Cambridge, UK: CUP.
- Boyd, R. 1991. “Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds.” *Philosophical Studies* 61: 127-48.
- Boyd, R. 1999. “Homeostasis, Species, and Higher Taxa.” In *Species: New Interdisciplinary Essays* (ed. Wilson). Cambridge MA: MIT Press.

- Boysen, S., Bernston, G, Hannan, M, & Cacciopo, J. 1996. "Quantity–based inference and symbolic representation in chimpanzees (Pan troglodytes)." *Journal of Experimental Psycholology: Animal Behaviour Processes*, 22/1: 76–86.
- Brandom, R. B. 1994. *Making it explicit*. Cambridge, MA: Harvard University Press.
- Brigandt, I. 2011. "Natural Kinds and Concepts: A Pragmatist and Methodologically Naturalist Account." In Knowles & Reidenfelt (eds.), *Pragmatism, Science and Naturalism.* Peter Lang.
- Brooks, R. 1991. "Intelligence without representation." *Artificial Intelligence,* 47: 139-59.
- Brownstein, M. 2016. "Implicit Bias." In Zalta (ed.), *Stanford Encyclopedia of Philosophy.*
- Bruner, B. 1990. *Acts of Meaning*. Cambridge, MA: HUP.
- Buckwalter, W. & Phelan, M. 2014. "Phenomenal Consciousness Disembodied". In Systma (ed.), *Advances in Experimental Philosophy of Mind*. Bloomsbury Academic.
- Butterfill, S. & Apperly, I. 2013. "How to Construct a Minimal Theory of Mind." *Mind & Language* 28/5: 606-37.
- Callaghan, T., Rochat, P., Lillard, A., Claux, M., Odden, H., Itakura, S., Tapanya, S., & Singh, S. 2005. "Synchrony in the onset of mental state reasoning." *Psychological Science, 16*: 378-384.
- Caprara, G. V. & Perugini, M. 1994. "Personality described by adjectives: the generalizability of the Big Five to the Italian lexical context." *European Journal of Personality* 8/5: 357-69.
- Carnap, R. 1959. "Beobachtungssprache und theoretische Sprache." *Logica: Studia Paul Bernays dedicta*. Neuchatel: Griffon.

- Carnap, R. 1961. "On the Use of Hilbert's e-Operator in Scientific Theories." In *Essays on the Foundations of Mathematics* (eds. Bar-Hillel *et al*). Jerusalem: Magnes Press.
- Carnap, R. 1963. "Replies and Systematic Expositions." In *The Philosophy of Rudolph Carnap* (ed. Schilpp). La Salle, IL: Open Court.
- Carnap, R. 1966. *Philosophical Foundations of Physics*. New York: Basic Books, 1966.
- Carruthers, P. 2006. *The Architecture of Mind*. Oxford: OUP.
- Carruthers, P. 2009. "How we know our own minds: the relationship between mindreading and metacognition." *Behavioural and Brain Science* 32/2: 121-38.
- Carruthers, P. 2011. *The Opacity of Mind.* Oxford, UK: OUP.
- Carruthers, P. 2013. "Mindreading in Infancy." *Mind and Language* 28/2: 141-72.
- Carruthers, P. Forthcoming. "Mindreading in adults: evaluating two-systems views." *Synthese.*
- Carruthers, P., Laurence, S. & Stich, S. 2005. *The Innate Mind.* Oxford, UK: OUP.
- Cattell, R. B. 1943. "The description of personality: basic traits resolved into clusters." *Journal of Abnormal and Social Psychology* 38: 476-506.
- Cattell, R. B. 1946. *The Description and Measurement of Personality*. Yonkers, NY: World Book.
- Cattell, R. B. 1947. "Confirmation and clarification of primary personality factors." *Psychometrics* 12: 197-220.
- Cattell, R. B. 1948. "The primary personality factors in women compared with those in men." British Journal of Psychology 1: 114-30.
- Chakravartty, A. 2015. "Scientific Realism." *Stanford Encyclopedia of Philosophy (Fall 2015 Edition).*

- Chartrand, T. & Bargh, J. 1999. "The chameleon effect: The perception-behavior link and social interaction." *Journal of Personality and Social Psychology* 76/6: 893–910.
- Chisholm, R. M. 1957. *Perceiving: A philosophical study.* Ithaca, NY: Cornell University Press.
- Choi, S. & Fara, M. 2016. "Dispositions." In Zalta (ed.), *The Stanford Encyclopedia of Philosophy.*
- Chomsky, N. 1959. "A Review of B. F. Skinner's *Verbal Behaviour*." *Language* 35/1: 26-58.
- Churchland, P.M. 1979. *Scientific realism and the plasticity of mind.* Cambridge, UK: CUP Press.
- Churchland, P.M. 1981. "Eliminative materialism and the propositional attitudes." *The Journal of Philosophy,* 78/2: 67-90.
- Clark, A. & Chalmers, D. 1998. "The Extended Mind." *Analysis,* 58/1: 7-19.
- Clark, A. 1987. "From Folk Psychology to Naive Psychology." *Cognitive Science* 11/2: 139-54.
- Clark, A. 1990. "Connectionism, competence, and explanation." *The British Journal for the Philosophy of Science* 41/2: 195-222.
- Clark, A. 2005. "Intrinsic content, active memory, and the extended mind." *Analysis* 65: 1-11.
- Clark, A. 2006. "Language, embodiment, and the cognitive niche." *TRENDS in Cognitive Science,* 10/8: 370-4.
- Clark, A. 2008. *Supersizing the Mind.* Oxford, UK: OUP.
- Clark, A. 2010. "Coupling, Constitution, and the Cognitive Kind." In Menary (ed.), *The Extended Mind*. Cambridge, MA: MIT Press.
- Clark, A. 2013. "Whatever Next?" *Behavioral and Brain Sciences,* 36/03: 181-24.
- Clark, A. 2016. *Surfing Uncertainty.* Oxford: OUP.
- Clark, A., & Prinz, J. Manuscript. "The Absence of Mind."

- Clark, A. & Millican, P. (eds.) 1999. *Connectionism, Concepts and Folk Psychology*. Oxford: OUP.
- Coleman, S. 2011. "There Is No Argument That The Mind Extends" *The Journal of Philosophy,* 108/2: 100-108.
- Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M., & Michel, F. 2000. "The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients." *Brain,* 123: 291–307.
- Colombo, M. Forthcoming. "Social motivation in computational neuroscience. Or if brains are prediction machines, then the Humean theory of motivation is false." In Kiverstein (ed.), *Routledge Handbook of Philosophy of the Social Mind*. Routledge.
- Cooper, R. 2004. "Why Hacking is wrong about human kinds." *British Journal for the Philosophy of Science* 55: 73-85.
- Cooper, R. 2013. "Natural Kinds." In Thornton, Graham, Sadler, & Davies (eds.), *The Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press.
- Craver, C. 2001. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science,* 68: 53-74.
- Craver, C. 2012. "Functions and Mechanisms: A Perspectivalist Account." In Huneman (ed.), *Functions*. Dordrecht: Springer.
- Craver, C. & Tabery, J. 2016. "Mechanisms in Science." In Zalta (ed.), *The Stanford Encyclopedia of Philosophy.*
- Cunningham, J. P. & Yu Byron, M. 2014. "Dimensionality reduction for large-scale neural recording." *Nature Neuroscience* 17/11: 1500-9.
- Danziger, E. 2011. "The Learning of Mind." *Suomen Anthropologi: Journal of the Finnish Anthropological Society*, 36/4: 51-3.
- Danziger, K. 2008. *Marking the Mind*. Cambridge, UK: CUP.

- Davidson, D. 1985. “Rational animals.” In LePore & McLaughlin (eds.), *Actions and events*. Oxford: Blackwell.
- Davies, M. & Stone, T. (eds.) 1995. *Folk Psychology.* Oxford, UK: Blackwell.
- De Bruin, L. & Newen, A. 2014. “The developmental paradox of false belief understanding.” 191/3: 297-320.
- De Jaegher, H. & Di Paolo, E. 2007. “Participatory Sense-Making: An enactive approach to social cognition.” *Phenomenology and the Cognitive Sciences*, 6(4): 485-507.
- Dennett, D. 1969. *Content and Consciousness*. London: Routledge & Keagan Paul.
- Dennett, D. 1978. “Beliefs about beliefs.” *Behavioral and Brain Sciences* 1: 568-70.
- Dennett, D. 1984. *Elbow Room.* Cambridge, MA: MIT Press.
- Dennett, D. 1987. *The Intentional Stance.* Cambridge, MA: HUP.
- Dennett, D. 2003. *Freedom Evolves*. New York: Viking Books.
- DeVries, W. 2006. “Folk psychology, theories, and the Sellarsian roots.” In Wolf & Lance (eds.), *The Self-Correcting Enterprise.* Amsterdam: Rodopi.
- Devitt, M. 2008. “Resurrecting Biological Essentialism.” *Philosophy of Science* 75: 344-82.
- Dewhurst, J. 2016. “Individuation Without Representation.” *British Journal for the Philosophy of Science.*
- Dewhurst, J. Forthcoming. “Folk Psychology and the Bayesian Brain.” In Metzinger & Wiese (eds.), *Philosophy and Predictive Processing.*
- Dewhurst, J. Manuscript. “Much Ado About Nothing.”
- Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. 1992. "Understanding motor events: a neurophysiological study". *Experimental Brain Research*, 91: 176–180.

- Digman, J. M. 1990. "Personality Structure: Emergence of the Five-Factor Model." *Annual Review of Psychology* 41: 417-40.
- Doherty, M. 2011. "A Two-Systems Theory of Social Cognition." In Roessler, Lerman, & Eilan (eds.), *Perception, Causation, and Objectivity.* Oxford, UK: OUP.
- Drayson, Z. 2012. "The Uses and Abuses of the Personal/Subpersonal." *Philosophical Perspectives* 26/1: 1-18.
- Drayson, Z. 2014. "The Personal/Subpersonal Distinction." *Philosophy Compass* 9/5: 338-46.
- Drobnak, F. T. 2009. "On the merits and shortcomings of semantic primes and natural semantic metalanguage in cross-cultural translation." *English Language Overseas Persepctives and Enquiries,* 6/1-2: 29-41.
- Dupré, J. 1981. "Natural Kinds and Biological Taxa." *The Philosophical Review* 90/1: 66-90.
- Dupré, J. 1995. *The Disorder of Things.* Cambridge, MA: HUP.
- Dupré, J. 1996. "Promiscuous Realism: A Reply to Wilson." *British Journal for the Philosophy of Science* 47: 441–444.
- Dupré, J. 1999. "Are Whales Fish?" In Medin & Atran (eds.), *Folkbiology*, pp. 461-76. Harvard, MA: MIT Press.
- Ekman, P. 1972. "Universals and cultural differences in facial expressions of emotion." In Cole (ed.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press.
- Ekman, P. 1999. "Basic emotions." In Dalgleish & Power (eds.), *Handbook of cognition and emotion*. New York: Wiley.
- Ereshefsky, M. & Reydon, T. 2015. "Scientific kinds." *Philosophical Studies* 172/4: 969-86.
- Feyerabend, P. 1963. "Mental events and the brain." *Journal of Philosophy,* 60: 295-6.

- Figdor, C. 2014. “On the proper domain of psychological predicates.” *Synthese*.
- Fodor, J. 1968. *Psychological Explanation.* New York: Random House.
- Fodor, J. 1975. *The Language of Thought.* Cambridge, MA: HUP.
- Fodor, J. 1987. *Psychosemantics.* Cambridge, MA: MIT Press.
- Fodor, J. & Pylyshyn, Z. 1981. “How direct is visual perception?” *Cognition* 9/2: 139-96.
- Fodor, J., & Pylyshyn, Z., 1988. “Connectionism and Cognitive Architecture: a Critical Analysis,” *Cognition*, 28: 3–71.
- Frege, G. 1892/1980. “On Sense and Reference.” In Geach & Black (eds. and trans.), *Translations from the Philosophical Writings of Gottlob Frege.* Oxford: Blackwell.
- Friston, K., & Frith, C. 2015. “A Duet for one.” *Consciousness and Cognition,* 36: 390-405.
- Friston, K., Mattout, J. & Kilner, J. 2011. “Action understanding and active inference.” *Biology Cybernetics* 104: 137-60.
- Fulkerson, M. 2014. “Rethinking the senses and their interactions: the case for sensory pluralism.” *Frontiers in Psychology* 5: 1426.
- Gallagher, S. 2008a. “Inference or interaction.” *Philosophical Explorations* 11/3: 163-74.
- Gallagher, S. 2008b. “Direct perception in the intersubjective context.” *Consciousness and Cognition* 17: 535-43.
- Gallagher, S. 2012. “Neurons, neonates and narrative: From embodied resonance to empathic understanding.” In Foolen, Lüdtke, Zlatev, & Racine (eds.), *Moving Ourselves, Moving Others*. Amsterdam: John Benjamins: 167-96.
- Garzón, F. 2008. “Towards a general theory of antirepresentationalism.” *The British Journal for the Philosophy of Science,* 59: 259-92.

- Geach, P. T. 1957. *Mental Acts: Their content and their objects.* London: Routledge and Kegan Paul.
- van Gelder, T. 1995. "What might cognition be, if not computation?" *The Journal of Philosophy,* 91: 345-81.
- Gelman, S. 2005. "Essentialism in Everyday Thought." *Psychological Science Agenda*. http://www.apa.org/science/about/psa/2005/05/gelman.aspx
- Gibson, J.J. 1966. *The Senses Considered as Perceptual Systems.* Boston: Houghton Mifflin.
- Gibson, J.J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
- Glennan, S. 1996. "Mechanisms and the Nature of Causation." *Erkenntnis,* 44: 49-71.
- Glennan, S. 2002. "Rethinking Mechanistic Explanation." *Philosophy of Science,* 69/3: S342-S353.
- Goldman, A. 1989. "Interpretation Psychologized." *Mind and Language* 4: 161-85.
- Goldman, A. 2006. *Simulating Minds.* Oxford, UK: OUP.
- Gopnik, A. & Wellman, H. 1992. "Why the Child's Theory of Mind Really is a Theory." *Mind and Language* 7: 145-72.
- Gordon, R. 1986. "Folk Psychology as Simulation." *Mind and Language* 1/2: 158-71.
- Hacking, I. 1982. "Wittgenstein the psychologist." *New York Review of Books* 29/5.
- Hacking, I. 1995. "The Looping Effects of Human Kinds." In Sperber, Premack, & Premack (eds.), *Causal Cognition, an Interdisciplinary Approach.* Oxford, UK: OUP.
- Hacking, I. 2000. *The Social Construction of What?* Cambridge, MA: HUP.
- Hacking, I. 2006. "Kinds of People." *Proceedings of the British Academy* 151: 285-318.

- Hacking, I. 2007. "Natural Kinds: Rosy Dawn, Scholastic Twilight." *Royal Institute of Philosophy Supplements* 61: 203-39.
- Hales, S. D. 2006. *Relativism and the foundations of philosophy.* Cambridge, MA: MIT Press.
- Hamilton, S. 2001. *Indian Philosophy: A Very Short Introduction*. Oxford: OUP.
- Harman, G. 1978. "Studying the chimpanzee's theory of mind." *Behavioural and Brain Sciences* 1: 576-7.
- Heal, J. 1986. "Replication and Functionalism". In *Language, Mind, and Logic* (ed. Butterfield). Cambridge: CUP.
- Heil, J. 2011. "The senses." In F. MacPherson (ed.) *The senses.* Oxford: OUP.
- Hendry, R. F. 2006. "Elements, Compounds, and Other Chemical Kinds." *Philosophy of Science* 73: 864-75.
- Henrich, J. 2009. "The evolution of costly displays, cooperation, and religion: Credibility enhancing displays and their implications for cultural evolution." *Evolution and Human Behavior* 30/4:244–60.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N., Hill, K., Gil-White, F., Gurven, M., Marlowe, F., Patton, J., & Tracer, D. 2005. "'Economic Man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies." *Behavioral and Brain Sciences* 28: 795–855.
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, A., Henrich, N., Lesorogol, C., Marlowe, M., Tracer, D., & Ziker, J. 2006. "Costly punishment across human societies." *Science* 312: 1767–769.
- Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C.,

Marlowe, M., Tracer, D., & Ziker, J. 2010. "Markets, religion, community size, and the evolution of fairness and punishment." *Science* 327: 1480–484.

- Henrich, J., Heine, S., & Norenzayan, A. 2010. "The weirdest people in the world?" *Behavioural and Brain Sciences,* 33/2-3: 61-83.
- Heyes, C. 1998. "Theory of mind in nonhuman primates." *Behavioral and Brain Sciences*, 21: 101–134.
- Hiatt, L. 1978. "Classification of the Emotions." In *Australian Aboriginal Concepts* (ed. Hiatt). Canberra: Australian Institute of Aboriginal Studies.
- Hickok, G. 2009. "Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans". *Journal of Cognitive Neuroscience*, 21/7: 1229-43.
- Hobson, J. & Friston, K. 2014. "Consciousness, dreams, and inference." *Journal of Consciousness Studies*, 21: 6-32.
- Hochstein, E. 2016a. "Categorizing the Mental." *The Philosophical Quarterly* 66/265: 745-59.
- Hochstein, E. 2016b. "One mechanism, many models." *Synthese*, 193/5: 1387-407.
- Hofstee, W. K. B., Kiers, H. A., de Raad, B., & Goldberg, L. R. 1997. "A comparison of Big-Five structures of personality traits in Dutch, English, and German." *European Journal of Personality*, 11: 15-31.
- Hood, B. 2004. "Is looking good enough or does it beggar belief?" *Developmental Science* 7/4: 415-7.
- Horner, V. & Whiten, A. 2005. "Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*)." *Animal Cognition* 8: 164–81.
- Howard, A. 1985. "Ethnopsychology and the Prospects for a Cultural Psychology." In White & Kirkpatrick (eds.), *Person, Self, and Experience*. Berkely: University of California Press.

- Hohwy, J. 2012. “Attention and conscious perception in the hypothesis testing brain.” *Frontiers in Psychology*.
- Hohwy, J. 2013. *The Predictive Mind.* Oxford: OUP.
- Huber, F. & Schmidt-Petri, C. 2016. *Degrees of Belief.* Springer: Synthese Library.
- Huebner, B., Bruno, M., & Sarkissian, H. 2010. "What Does the Nation of China Think about Phenomenal States?" *Review of Philosophy and Psychology*, 1/2: 225-243.
- Hurley, S. 2010. “The Varieties of Externalism.” In Menary (ed.), *The Extended Mind*. Cambridge, MA: MIT Press.
- Hutto, D. 2008. *Folk Psychological Narratives.* Cambridge, MA: MIT Press.
- Hutto, D. 2009. “Lessons from Wittgenstein: elucidating folk psychology.” *New Ideas in Psychology*, 27: 118-32.
- Hutto, D. 2016. “Basic social cognition without mindreading: minding minds without attributing content.” *Synthese*.
- Hutto, D., Southgate, V., & Schwenkler, J. (eds.) 2011. “Social Cognition: Mindreading and Alternatives”. Special issue of *Review of Philosophy and Psychology*.
- Irvine, E. 2013. *Consciousness as a Scientific Concept.* Springer.
- Jenson, J. C. 2016. “The Belief Illusion.” *The British Journal for the Philosophy of Science*, 67: 965-995.
- John, O. & Srivastava, S. 1999. “The Big-Five Trait Taxonomy.” In Pervin & John (eds.), *Handbook of Personality: Theory and Research (2nd ed.).* New York: Guildford.
- Kendig, C (ed.). 2015. *Natural Kinds and Classification in Scientific Practice.* Routledge.
- Khalidi, M. A. 2013. *Natural Categories and Human Kinds.* Cambridge, UK: CUP.

- Kim, A. "Wilhelm Maximilian Wundt." *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*.
- Klein, C. 2012. "Cognitive Ontology and Region- versus Network-Oriented Analyses." *Philosophy of Science,* 79/5: 952-60.
- Knobe, J. & Prinz, J. 2008. "Intuitions bout consciousness: Experimental studies." *Phenomenology and the Cognitive Sciences,* 7: 67-83.
- Kornblith, H. 1993. *Inductive Inference and Its Natural Ground.* Cambridge, MA: MIT Press.
- Kripke, S. 1980. *Naming and Necessity.* Cambridge, MA: HUP.
- Klages, L. 1926. *The Science of Character*. (Translated 1932). London: Allen & Unwin
- Laland, K. N., Odling-Smee, F. J., & Feldman, M. W. 1999. "Evolutionary consequences of niche construction and their implications for ecology." *Proceedings of the National Academy of Sciences*, 96/18: 10242-7.
- Laland, K. N., Odling-Smee, F. J., & Feldman, M. W. 2001. "Niche construction, ecological inheritance, and cycles of contingency in evolution." In Oyama, Griffiths, & Gray (eds.), *Cycles of contingency.* Cambridge, MA: MIT Press.
- Lavelle, J.S. 2012. "Theory-theory and the Direct Perception of Mental States." *Review of Philosophy and Psychology,* 3/2: 213-30.
- Lavelle, J.S. 2016. "Cross-Cultural Considerations in Social Cognition." *A Handbook to Social Cognition* (ed. Kiverstein). Routledge.
- Leach, S. & Tartaglia, J. (eds.) 2014. *Mind, Language, and Metaphilosophy.* Cambridge, UK: CUP.
- Lebra, T. S. 1993. "Culture, self, and communication in Japan and the United States." In W. B. Gudykunst (ed.), *Communication in Japan and the United States.*
- Lenartowicz, A., Kalar, D.J., Congdon, E., & Poldrack, R.A. 2010. "Towards an ontology of cognitive control." *Topics in Cognitive Science*, 2:

678–92.

- Leslie, A.M. 1994. "ToMM, ToBy, and agency: Core architecture and domain specificity." In L. Hirschfeld and S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge, MA, Cambridge University Press: 119–148.
- Leslie, A. M. 2000. "'Theory of mind' as a mechanism of selective attention." In M. Gazzaniga (ed.), *The New Cognitive Neurosciences*. Cambridge, MA, MIT Press: 1235–1247.
- Levin, J. 2016. "Functionalism." *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition).*
- LeVine, R. (ed.) 2010. *Psychological Anthropology.* Oxford, UK: Wiley Blackwell.
- Lewis, D. 1966. "An Argument for the Identity Theory." *Journal of Philosophy*, 63: 17–25.
- Lewis, D. 1970. "How to Define Theoretical Terms." *Journal of Philosophy*, 67: 427–46.
- Lewis, D. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50: 249–58.
- Lewis, D. 1980. "Mad Pain and Martian Pain." *Readings in the Philosophy of Psychology* (ed. Block): 216-222. Cambridge, MA: HUP.
- Lillard, A. 1998. "Ethnopsychologies." *Psychological Bulletin*, 123/1: 3-32.
- Liu, D., Wellman, H., & Tardif, T. 2008. "Theory of Mind Development in Chinese Children." *Developmental Psychology* 44/2: 523-31.
- Ludwig, D. 2015. "Indigenous and Scientific Kinds." *British Journal for the Philosophy of Science,* available online: doi: 10.1093/bjps/axv031
- Luhrmann, T. 2011. "Toward An Anthropological Theory of Mind." *Suomen Anthropologi: Journal of the Finnish Anthropological Society*, 36/4.
- Lycan, W. 1988. *Judgement and Justification.* Cambridge: CUP.

- McGeer, V. 2007. "The regulative dimension of folk psychology." In Hutto & Ratcliffe (eds.), *Folk Psychology Reassessed.*
- Machery, E. 2005. "Concepts are not a natural kind." *Philosophy of Science* 72/3: 444-67.
- Machery, E. 2009. *Doing Without Concepts.* Oxford: OUP.
- Machery, E. 2014. "Concepts: Investigating the Heterogeneity Hypothesis." In Systma (ed.), *Advances in Experimental Philosophy of Mind*. Bloomsbury Academic.
- Machery, E. Forthcoming. *Philosophy within its Proper Bounds.* Oxford: OUP.
- Macpherson, F. 2011. "Taxonomising the Senses." *Philosophical Studies* 153: 123-42.
- Magnus, P. D. 2012. *Scientific Enquiry and Natural Kinds.* Palgrave Macmillan.
- Mameli, M. 2001. "Mindreading, mindshaping, and evolution." *Biology and Philosophy* 16/5: 595-626.
- Martin, A., & Chao, L. L. 2001. "Semantic memory and the brain: Structure and processes." *Current Opinions in Neurobiology,* 11: 194–201.
- Matthews, R. 2013. "Belief and Belief's Penumbra." In *New Essays on Belief* (ed. Nottelmann). UK: Palgrave Macmillan.
- Matthews, R. Unpublished draft. *The Cultural Construction of Belief.*
- McCaffrey, J. 2015. "The Brain's Heterogeneous Functional Landscape." *Philosophy of Science* 82/5: 1010-22.
- McGeer, V. 2007. "The Regulative Dimension of Folk Psychology." In Hutto & Ratcliffe (eds.), *Folk Psychology Re-assessed.* Springer.
- Menary, R. (ed.) 2010. *The Extended Mind.* Cambridge, MA: MIT Press.
- Michel, J. B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, the Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, & E. L. Aiden. 2011. "Quantitative Analysis of

Culture Using Millions of Digitized Books." *Science* 331/6014: 176–82.

- Millikan, R. 1996. "Pushmi-Pullyu Representations." *Philosophical Perspectives*, 9: 185-200.
- Millikan, R. 1999. "Historical Kinds and the Special Sciences." *Philosophical Studies* 95: 45–65.
- Morton, A. 1980. *Frames of Mind: Constraints on the Common-sense Conception of the Mental.* Oxford: Clarendon Press.
- Morton, A. 2007. "Folk psychology does not exist." In Hutto & Ratcliffe (eds.), *Folk Psychology Re-assessed.* Springer.
- Montandon, A. 2004. "Tangled in the Machine." MSc Thesis, University of Plymouth.
- Nado, J. 2014. "The Role of Intuition." In Systma (ed.), *Advances in Experimental Philosophy of Mind*. Bloomsbury Academic.
- Nagel, S.K., Carl, C., Kringe, T., Märtin, R., & König, P. 2005. "Beyond sensory substitution – learning the sixth sense". *Journal of neural engineering* 2/4: 13–26.
- Naito, M. & Koyama, K. 2006. "The Development of False Belief Understanding in Japanese Children: Delay or Difference?" *International Journal of Behavioural Development.*
- Newton, A. M., & de Villiers, J. G. 2007 "Thinking while talking: adults fail nonverbal false-belief reasoning." *Psychological Science* 18/7: 574-9.
- Norman, W. T. 1963. "Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings." *Journal of Abnormal and Social Psychology* 66: 574-83.
- Nudds, M. 2004. "The significance of the senses." *Proceedings of the Aristotelian Society* 104/1: 31–51.
- O'Brien, G. J. 1991. "Is connectionism commonsense?" *Philosophical Psychology* 4: 165-78.

- Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. 2003. *Niche Construction: the neglected process in evolution.* Princeton University Press.
- Ortony, A. & Turner, T. J. 1990. "What's Basic About Basic Emotions?" *Psychological Review* 97/3: 315-31.
- Onishi, K. & Baillergeon, R. 2005 "Do 15-Month-Old Infants Understand False Beliefs?" *Science* 308/5719: 255-8.
- Perner, J. & Howes, D. 1992. "'He Thinks He Knows': And More Developmental Evidence Against the Simulation (Role Taking) Theory." *Mind & Language*, 7: 72–86.
- Pettigrew, R. 2015. "Pluralism about Belief States." *Aristotelian Society Supplementary Volume,* 89: 187-204.
- Piccinini, G. 2004a. "Functionalism, Computationalism, and Mental Contents." *Canadian Journal of Philosophy* 34/3: 375-410.
- Piccinini, G. 2004b. "Functionalism, Computationalism, and Mental Contents." *Studies in the History and Philosophy of Science* 35: 811-33.
- Piccinini, G. & Craver, C. 2011. "Integrating psychology and neuroscience: functional analyses as mechanism sketches." *Synthese* 183: 283-311.
- Pickering, M. & Chatter, N. 1995. "Why Cognitive Science Is Not Formalized Folk Psychology." *Minds and Machines* 5: 309-37.
- Poldrack, R. 2006. "Can cognitive processes be inferred from neuroimaging data?" *TRENDS in Cognitive Science,* 10/2: 59-63.
- Poldrack, R. 2010. "Mapping Mental Function to Brain Structure." *Perspectives on Psychological Science,* 5/6: 753-61.
- Premack, D. & Woodruff, G. 1978. "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences* 4: 515-26.
- Price, C. & Friston, K. 2005. "Functional ontologies for cognition." *Cognitive Neuropsychology*, 22/3: 262-75.
- Povinelli, D. & Vonk, J. 2003. "Chimpanzee minds: suspiciously human?" *TRENDS in Cognitive Science*, 7/4: 157-10.

- Putnam, H. 1960. “Minds and machines.” *Journal of Symbolic Logic*: 57-80.
- Putnam, H. 1967. “Psychological Predicates”. In *Art, Philosophy, and Religion*. Pittsburgh, PA: University of Pittsburgh Press.
- Putnam, H. 1975. *Mathematics, Matter and Method.* Cambridge, UK: CUP.
- Quadt, L. Forthcoming. “Action-Oriented Predictive Processing and Social Cognition.” In Metzinger & Wiese (eds.), *Philosophy and Predictive Processing.*
- Quine, W. V. O. 1969. “Natural Kinds.” In *Ontological Relativity and Other Essays.*
- Hutto, D. & Ratcliffe, M. 2007. *Folk Psychology Re-assessed.* Springer.
- Ramsey, F. 1931. “Theories”. *The Foundation of Mathematics* (ed. Braithwaite). London: Routledge & Kegan Paul.
- Ramsey, W., Stich, S. & Garon, J. 1991. “Connectionism, eliminativism, and the future of folk psychology.” In Ramsey, Stich & Rumelhart (eds.), *Philosophy and Connectionist Theory*. Lawrence Erlbaum.
- Ratcliffe, M. 2007. “Why folk psychology is not folk psychology.” *Phenomenology and the Cognitive Sciences,* 5: 31-52.
- Ratcliffe, M. 2009. “There Are No Folk Psychological Narratives.” *Journal of Consciousness Studies* 16/6-8: 379-406.
- Ravenscroft, I. 2010. "Folk Psychology as a Theory." *The Stanford Encyclopedia of Philosophy*.
- Regier, T. & Kay, P. “Language, thought, and color: Whorf was half right.” *Trends in Cognitive Science* 795: doi:10.1016/j.tics.2009.07.001
- Reuter, K., Phillips, D., & Systma, J. 2014. “Hallucinating Pain.” In Systma (ed.), *Advances in Experimental Philosophy of Mind*. Bloomsbury Academic.
- Richards, G. 2000. “Britain on the Couch: The Popularisation of Psychoanalysis in Britain 1918-1940.” S*cience in Context,* 13/2: 183-230.

- Ritchie, I. 1991. “Fusion of the faculties.” In Howes (ed.), *The varieties of sensory experience*. Toronto, Canada: UTP. 192-202.
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. 1996. "Premotor cortex and the recognition of motor actions." *Cognitive Brain Research*. 3 (2): 131–141
- Robbins, J. & Rumsey, A. 2008. “Cultural and Linguistic Anthropology and the Opacity of Other Minds.” *Anthropological Quarterly,* 81/2: 407-20.
- Rorty, R. 1965. “Mind-Body Identity, Privacy, and Categories.” *The Review of Metaphysics,* 19/1: 24-54.
- Ross, D. & Ladyman, J. 2010. “The Alleged Coupling-Constitution Fallacy and the Mature Sciences.” In Menary (ed.), *The Extended Mind*. Cambridge, MA: MIT Press.
- Rozin, P. “’Taste-smell confusions’ and the duality of the olfactory sense.” *Percept Psychophysics* 31: 397- 401.
- Rumelhart, D. E. & McClelland, J. L. 1986. *Parallel Distributed Processing.* Cambridge, MA: MIT Press.
- Rupert, R. 2004. “Challenges to the hypothesis of extended cognition.” *Journal of Philosophy* 101: 389-428.
- Rupert, R. 2013. “Memory, Natural Kinds, and Cognitive Extension; or, Martians Don’t Remember, and Cognitive Science Is Not About Cognition.” *Review of Philosophy and Psychology*, 4: 25-47.
- Rupert, R. Forthcoming. “Individual Minds as Groups, Group Minds as Individuals.” In Kaldis (ed.), *Mind and Society: Cognitive Science Meets the Philosophy of the Social Sciences*. Synthese Philosophy Library Studies in Epistemology, Logic, Methodology, & Philosophy of Science.
- Russell, B. 1910. “Knowledge by acquaintance and knowledge by description.” *Proceedings of the Aristotelian Society*, 11: 108-28.
- Ryle, G. 1949. *The Concept of Mind.* University of Chicago Press.

- Sabb, F. W., Bearden, C. E., Glahn, D. C., Parker, D. S., Freimer, N., & Bilder, R. M. 2008. "A collaborative knowledge base for cognitive phenomics." *Molecular Psychiatry*, 13/4: 350–60.
- Schwitzgebel, E. 2002. "A phenomenal, dispositional account of belief." *Noûs*, 36/2: 249–75.
- Schwitzgebel, E. 2015. "Belief." In Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.
- Segall, M., Campbell, D., & Herskovits, M. 1966. *The Influence Of Culture On Visual Perception*. Oxford, UK: Bobbs-Merrill.
- Sellars, W. 1956. "Empiricism and the Philosophy of Mind." *Minnesota Studies in the Philosophy of Science.* References refer to the 1997 reprint published by Harvard University Press.
- Sellars, W. 1963. *Science, Perception and Reality.* London: Routledge & Keegan Paul.
- Shagrir, O. 2001. "Content, Computation and Externalism," *Mind*, 110(438): 369–400.
- Shapiro, L. 2010. *Embodied Cognition.* Routledge.
- Sinclair, S., Huntsinger, J., Skorinko, J., & Hardin, C. D. 2005. "Social tuning of the self: Consequences for the self evaluations of stereotype targets." *Journal of Personality and Social Psychology,* 89/2: 160-75.
- Slater, M. 2015. "Natural Kindness." *British Journal for Philosophy of Science* 66: 375-411.
- Smart, J. J. C. 1959. "Sensations and brain processes." *Philosophical Review* 68: 141-56.
- Smith, G. M. 1967. "Usefulness o f peer ratings of personality in educational research." *Educational and Psychological Measurement* 27: 967-84.
- Sosa, E. 2009. "A defense of the use of intuitions in philosophy." In Murphy & Bishop (eds.), *Stich and His Critics*. Wiley-Blackwell.

- Southgate, V., Senju, A., & Csibra, G. 2007. "Action anticipation through attribution of false belief by two-year-olds." *Psychological Science*, 18(7): 587–92.
- Southgate, V., & Vernetti, A. 2014. "Belief-based prediction in pre-verbal infants." *Cognition* 130: 1-10.
- Spratling, M. W. "A review of predictive coding algorithms." *Brain and Cognition*.
- Sprevak, M. 2009. "Extended cognition and functionalism." *Journal of Philosophy* 106/9: 503-27.
- Stanford, K. "Underdetermination of Scientific Theory." In Zalta (ed.), *The Stanford Encyclopedia of Philosophy.*
- Sterelny, K. 2007. "Social intelligence, human intelligence, and niche construction." *Philosophical Transactions of the Royal Society of London: Series B*, 362: 719–30.
- Sterelny, K. 2012. *The evolved apprentice*. Cambridge, MA: MIT Press.
- Sterelny, K. 2015. "Content, Control, and Display." *Philosophia* 43/3: 549-64.
- Stich, S. 1978. "Belief and subdoxastic states." *Philosophy of Science* 45/4: 499-518.
- Stich, S. 1983. *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.
- Stich, S. & Nichols, S. 1992. "Folk Psychology: Simulation or Tacit Theory?" *Mind and Language*, 7: 35–71.
- Stich, S. & Nichols, S. 2003. "Folk psychology." In Stich & Warfield (eds.), *the Blackwell Guide to Philosophy of Mind*. Oxford: Basil Blackwell.
- Stich, S. & Ravenscroft, I. 1994. "What *is* folk psychology?" *Cognition,* 50: 447-68.
- Sullivan, J. 2016. "Construct Stabilization and the Unity of the Mind-Brain Sciences." *Philosophy of Science* 83: 662-673.

- Surian, L., Caldi, S. & Sperber, D. 2007. "Attribution of Beliefs by 13-Month Old Infants." Psychological Science 18/7: 580-6.
- Sutton, J. 2010. "Exograms and Interdisciplinarity." In Menary (ed.), *The Extended Mind*. Cambridge, MA: MIT Press.
- Systma, J. 2014. *Advances in Experimental Philosophy of Mind.* Bloomsbury Academic.
- Szirmak, Z. & De Raad, B. 1994. "Taxonomy and structure of Hungarian personality traits." *European Journal of Personality* 8: 95-117.
- Thagard, P. 2012. "Cognitive Science." *The Stanford Encyclopedia of Philosophy*.
- Thaler, L. & Goodale, M.A. 2016. "Echolocation in humans: an overview." *Wiley Interdisciplinary Reviews: Cognitive Science* 7/6: 382-93.
- Thomson, E., Carra, R., & Nicolelis, M. 2013. "Perceiving Invisible Light through a Somatosensory Cortical Prosthesis." *Nature Communications* 4: 1482.
- Tierney, H., Howard, C., Kumar, V., Kvaran, T., & Nichols, S. "How Many of Us Are There?" In Systma (ed.), *Advances in Experimental Philosophy of Mind*. Bloomsbury Academic.
- Trevarthen, C. 1979. "Communication and cooperation in early infancy: A description of primary intersubjectivity." In Bullowa (ed.), *Before Speech*. Cambridge: Cambridge University Press
- Trevarthen, C. & Hubley, P. 1978. "Secondary intersubjectivity: confidence, confiding and acts of meaning in the first year." In Lock (Ed.) *Action, gesture and symbol: The emergence of language*, London: Academic Press: 183-229.
- Turner, R. 2012. "The need for a systematic ethnopsychology." *Anthropological Theory,* 12/1: 29-42.
- Vinden, P. 1996. "Junin Quechua children's understanding of mind." *Child Development* 67: 1707-16.

- Weinberg, J., Nichols, S. & Stich, S. 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics* 29/1-2: 429-460.
- Weinberg, J., Gonnerman, C., Buckner, C., and Alexander, J. 2010. "Are philosophers expert intuiters?" *Philosophical Psychology*, 23: 331-355.
- Wellman, H. 2014. *Making Minds: How Theory of Mind Develops*. Oxford: OUP.
- Wellman, H., David, C., & Watson, J. 2001. "Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief." *Child Development,* 72/3: 655-84.
- Wheeler, M. 2010. "In Defense of Extended Functionalism." In Menary (ed.), *The Extended Mind*. Cambridge, MA: MIT Press.
- White, G. 1992. "Ethnopsychology." In *New Directions in Psychological Anthropology*, eds. Schwartz, White, & Lutz. Cambridge: CUP.
- Wierzbicka, A. 1986. "Human Emotions: Universal or Culture Specific?" *American Anthropologist* 88: 584-94.
- Wierzbicka, A. 1992. *Semantics, Culture, and Cognition: Universal Human Concepts in Culture.* Oxford: OUP.
- Wierzbicka, A. 2005. "Empirical Universals of Language as a Basis for the Study of Other Human Universals and as a Tool for Exploring Cross-Cultural Differences." *Ethos* 33/2: 256-91.
- Wierzbicka, A. 2010. "Defining Emotion Concepts." *Cognitive Science* 16/4: 539-81.
- Wikforss, A. 2010. "Are Natural Kind Terms Special?" In Beebee & Sabberton-Leary (eds.), *The semantics and metaphysics of natural kinds*. London: Routledge.
- Williamson, T. 2007. *The Philosophy of Philosophy.* Oxford: Blackwell.
- Wilkes, K. V. 1991. "The relationship between common-sense psychology and scientific psychology." *Synthese* 89: 15-39.

- Wimmer, H. & Perner, J. 1983. "Beliefs about beliefs." *Cognition* 13: 103-28.
- Woodward, J. 2014. "Scientific Explanation." In Zalta (ed.), *The Stanford Encyclopedia of Philosophy.*
- Wundt, W. 1912/1916. *Elements of Folk Psychology.* Schaub (trans.). London: G. Allen & Unwin.
- Zanna, M. P. & Pack, S. J. 1975. "On the self-fulfilling nature of apparent sex differences in behavior." *Journal of Experimental Social Psychology,* 11: 583-91.
- Zawidzki, T. 2008. "The function of folk psychology: mind reading or mind shaping?" *Philosophical Explorations* 11/3: 193-210.
- Zawidzki, T. 2013. *Mindshaping.* Cambridge, MA: MIT Press.