BELIEF AND COGNITIVE SCIENCE: THE CASE FOR MODEST INTEGRATIONISM

-

ANDREW GARFORD MOORE

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

Doctor of Philosophy

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

January 2021

Abstract

This study comprises a defence of **modest integrationism** towards the concept of belief. Modest integrationism holds that the concept of belief should be integrated into cognitive science. I describe the position as 'modest' because I accept that the concepts of cognitive science should be constantly refined in light of new empirical evidence, and that these refinements may ultimately result in the fracturing of the concept of belief, rendering it redundant and open to elimination. However, the current state of empirical evidence does not warrant elimination nor is elimination inevitable in the future.

I develop a framework to assess if elimination is appropriate based on whether a concept has scientific utility. I show that the concept of belief that we need to assess is not the multi-use concept that is simply taken from everyday English discourse. Rather, it is a doubly theoretical concept that 1) forms a key part of the core theory underlying our mental understanding abilities and 2) has been abstracted and reconstructed from the complexity of usage of the term 'belief' and other belief-like terms. I argue in favour of representationalism, demonstrating that belief possession is based on internal factors, and I reject the recent re-emergence of dispositionalism.

I focus on the case study of a debate within developmental psychology about the nature of infants' mental understanding abilities. I argue that the concept of belief plays a direct role in theories explaining empirical results, demonstrating that the concept still currently has scientific utility. Finally, I review empirical evidence that shows that proposed fracture lines within the concept of belief are not as clear cut as has been suggested and that infant metarepresentations are not easily categorised into sub-belief states. This suggests that the elimination of the concept of belief in the future is not a foregone conclusion.

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not previously been presented for an award at this, or any other, university.

Contents

Abstract 2

Chapter outlines 5

Introduction 7

1. Eliminativism: The Case Against Belief 10

  1.1 Starting assumptions – folk psychology and cognitive science 10

  1.2 Jenson's challenge 11

  1.3 Robustness and fragility 12

  1.4 What is a belief? 14

  1.5 Evidence of fragility? 15

  1.6 Is belief fragile? 17

  1.7 A case study – innateness in cognitive science 18

  1.8 A new framework for elimination 21

2. Folk 'belief' and philosophers' BELIEF 25

  2.1 constraints on methodology – is BELIEF pancultural? 27

  2.2 'Belief', BELIEF and the folk 28

    2.21 'Belief' as it functions in language 28

    2.22 BELIEF is often left unverbalised 33

    2.23 BELIEF is often used in combination with other concepts 35

  2.3 Philosophers' BELIEF 39

3. What is a *belief*? 46

  3.1 BELIEF and representationalism 46

  3.2 Dispositionalism 48

  3.3 Evaluating the positions 53

    3.31 The advantages of 'deep' accounts 53

    3.32 Are there advantages to the dispositionalist account? 55

4. BELIEF in Cognitive Science 60

  4.1 Direct and indirect concepts 61

  4.2 General concepts in science 62

  4.3 Infants mental understanding: rich or lean? 64

    4.31 A review of the empirical literature 65

    4.32 Interpretations of the data 69

    4.33 Rich or lean? 74

5. Beliefs about beliefs 76

  5.1 BELIEF as a direct concept in developmental psychology 77

  5.2 Doxastic and subdoxastic: division or continuum? 80

5.21 Proposed fractures in BELIEF 80
5.22 A spectrum of integration? 82
5.23 Problem states – infants' metarepresentations 87
Conclusion 90
Bibliography 93

Chapter outlines

1. Eliminativism: The Case Against Belief

In this chapter I consider an eliminativist challenge to the concept of belief. I reject the specific challenge but argue that belief is at threat of elimination. I develop a new framework to assess the threat.

1.1 **Starting assumptions – folk psychology and cognitive science.** Contrasting two types of eliminativism and introducing integrationism.

1.2 **Jenson's challenge.** The concept of belief should be eliminated because it is 'fragile'.

1.3 **Robustness and fragility.** Setting out robustness/fragility test for elimination.

1.4 **What is a belief?** Identifying a target account of belief that can be used in the test.

1.5 **Evidence of fragility?** A review of evidence suggesting that belief is fragile.

1.6 **Is belief fragile?** A critique of the robustness/fragility framework as a way of assessing whether belief should be eliminated.

1.7 **A case study – innateness in cognitive science.** A case study of a similar debate on the elimination of 'innateness' from cognitive science. Identification of lessons that we should learn for the present question.

1.8 **A new framework.** Setting out a new framework to determine whether belief should be eliminated. Identifying two types elimination.

2. Folk 'belief' and philosophers' BELIEF

In this chapter I show that the concept that is integrated into cognitive science is not belief as it is used in everyday language, but rather a doubly theoretical generalised concept of belief.

2.1 **Constraints on methodology – is BELIEF pancultural?** While this inquiry focuses on the usage of 'belief' in the English language, this section considers reasons that BELIEF may be pancultural.

2.2 **'Belief', BELIEF, and the folk.** How the term 'belief' functions in everyday English usage and the obstacles that this presents to those trying to abstract a generalised account of belief.

2.3 **Philosopher's BELIEF.** An account of the concept that philosophers have abstracted from everyday usage and a statement of the functional role of BELIEF.

3. What is a *belief*?

In this chapter I consider two accounts of what it is to have a *belief* – representationalism and dispositionalism. Adopting the dispositionalist account may provide a defence against the eliminativist threat. The dispositionalist account is rejected in favour of the representationalist.

3.1 **BELIEF and representationalism.** How the generalised account of the BELIEF fits with a representationalist account of *belief*.

3.2 **Dispositionalism**. An account of dispositionalism in its traditional form and the classic objections to it. Followed by a sympathetic description a more recent account of dispositionalism and how it attempts to avoid those objections.

3.3 **Evaluating the positions.** An evaluation of the two positions. A demonstration of the advantages of the representationalist account and the rejection of two proposed arguments in favour of the dispositionalist account.

4. BELIEF in cognitive science

In this chapter I set out the case study of the rich versus lean debate in developmental psychology – are infants' mental understanding abilities achieved by metarepresentations or behavioural rules?

4.1 **Direct and indirect concepts.** An account of two roles that concepts can play in scientific theories. I argue that only concepts that play a direct role can be said to have scientific utility.

4.2 **General concepts in science.** A demonstration that even concepts that can be reduced to lower-level concepts can play an important role in scientific theories in some domains.

4.3 **Infants mental understanding: rich or lean?** A detailed description of the rich versus lean debate in developmental psychology. Including a review of the empirical literature and expositions of how each of the different account explain the data. I conclude that the rich theory is to be preferred because it has proven more fruitful.

5. Beliefs about beliefs

In this chapter I argue that the concept of belief is playing a direct role in rich accounts of infants' mental understanding abilities. I also review empirical evidence to show that drawing distinctions within the category of belief is not straightforward.

5.1 **BELIEF as a direct concept in developmental psychology.** An identification of the concept of belief with technical terms used in developmental psychology such as 'attribute'.

5.2 **Doxastic and subdoxastic: division or continuum?** A review of the suggestion that BELIEF fractures into states that are inferentially integrated and those that are inferentially insulated. I review empirical evidence which suggests that inferential integration is a spectrum rather than a binary. Finally, I review the latest evidence on the cognitive role of infants metarepresentations and suggest that they are difficult to categorise into the proposed successor states.

Introduction

I'm in the office gossiping and I see Alison charging past in a hurry. I say to my colleague Sarah. "Why is Alison always charging around like that, why doesn't she just leave for her meeting earlier?", Sarah replies, "she thinks it makes her look busy".

Adult humans have an undoubted understanding of their own mind and the minds of others. We use these abilities constantly during our social lives to guide our behaviour and our interactions with others. We predict and explain behaviour, we can aim to please by anticipating and fulfilling others' desires, we can infer what people know from observing their perceptual input, we can even attempt to mislead by manipulating what they perceive. This nest of social abilities I will refer to – as neutrally as possible – as **mental understanding abilities**[1].

Broadly characterised, our mental understanding abilities allow us to take an agent's **input** (primarily their perceptions), combine this with any current knowledge we have of their mental states (if any) and use this to predict the agent's **output** – a wide-ranging category that includes overt behaviour, cognitive outputs such as decisions, inferences or judgements, and phenomenal outputs such as experiencing emotions, mood changes etc.[2]

When we use our mental understanding abilities, we will often make reference to a variety of different mental states. Why does Alison always rush around like she is constantly in a hurry? Because she **thinks** it will make her look busy, and Alison **wants** to look like she is always busy. Why do we expect that Hans will start loudly boasting about his achievements when Celeste enters the room? Because we know that he **desires** Celeste's affections, and he **believes** that if she hears of his achievements she will be impressed.

A common explanation of our mental understanding abilities is that we have a theory-like body of information about how these mental states interact with input and each other to produce output. This body of information is commonly known as folk psychology[3].

Eliminative materialism is a current of thought that has become popular over the past few decades. The most radical proponents of eliminativism (e.g. Paul Churchland, 1981) claim that folk psychology is an incorrect theory. States like beliefs, desires and intentions are the posits of this incorrect theory. As an incorrect theory, folk psychology will be replaced by a mature neuroscientific theory of the human brain and the posits of folk psychology will be eliminated. For the radical eliminativists, the elimination of beliefs and desires will not be limited to scientific discourse, it will also impact on "…our mutual understanding and even our introspection" (Churchland, 1981, p.67).

A more conservative form of eliminativism (e.g. Jenson, 2016) accepts the initial integration of the categories of folk psychology into cognitive science but argues that recent empirical data has shown that they are not precise enough for scientific purposes. The folk theory needs to be refined then, and as part of this its categories may be fractured into sub-states, leaving the old general terms redundant. Left without scientific utility, the general terms of folk psychology should be eliminated from scientific discourse. While not precise enough for scientific purposes, folk psychology still has a place and could enjoy its scientific retirement by continuing to play an important role in folk discourse.

[1] One common term for this nest of abilities is 'mindreading' e.g. Nichols and Stich (2003). However, I do not like the supernatural connotations of this term. These can be especially troublesome when communicating with those not familiar with the field.

[2] A characterisation inspired by Saxe (2005).

[3] See, for example, Fodor (1987).

My thesis is intended as a counterweight to this more conservative form of eliminativism. Focusing specifically on belief, I show that this concept does still play an important role in cognitive science and is likely to continue to do so for the foreseeable future. This position I label as 'modest integrationism'.

Following the terminology of Frankish (2004, p.1-2) the position is 'integrationist' because it holds that the concepts of folk psychology should be integrated into the theories of cognitive science. The position is 'modest' because this is not an argument that the concept of belief in cognitive science must remain fixed forever as it is used in folk discourse. Indeed, I will show how the concept of belief that has been integrated into cognitive science is not a crude and rough-edged concept that is to be straightforwardly lifted from everyday usage. Rather, abstracting from the complexity of everyday usage of 'belief' and other folk psychological terms, philosophers have identified the core concept that underlies our everyday usage of these terms.

'Modest' also because we must accept that this concept should be open to refinement as our scientific knowledge advances. One possible refinement is the fracturing of the concept of belief into successor states, which may allow scientists to develop more detailed theories that better match the data. However, I will demonstrate that the eliminativists are at best premature in heralding the redundancy of belief.

Firstly, no elimination can take place until scientists have fully developed successor states. I will review empirical data which suggests that drawing distinctions within the category of belief is not as straightforward as it might seem.

Secondly, even if such successor states were fully developed, cognitive science is an interdisciplinary endeavour. While a general concept of belief might be considered too imprecise in some disciplines, it may still play a useful role in others. I will demonstrate this by exploring a recent debate in cognitive science in which the concept of belief, albeit with a different label, plays a key role.

Before we start, a point on my usage of the word 'belief'. In a thesis such as this the word can be used in a variety of ways. We could be talking about the word 'belief' itself as it is mentioned in the previous sentence. We could be talking about what the word 'belief' refers to – be it a state, cluster of dispositions, or an abstract object. We could also be talking about the concept of belief, something I will argue has been abstracted from everyday usage.

To address this we can introduce notation, for example following Samuels (2007) we can refer to the concept of belief in small capitals (BELIEF), use inverted commas for the word 'belief' and other belief-like words (e.g. 'belief', 'believe', 'think that', etc.) and use lower case italics to refer to whatever it is that 'belief' refers to (*belief*).

Such a notation has limitations. Particularly with regards to what it is that 'belief' refers to – *belief* in my notation. The notational device doesn't distinguish between type and token, or the content of a belief and the vehicle of belief. If a schoolchild and a physicist have a belief with the content "the tides are caused by the moon's gravity" do they have the same *belief*? In certain respects, they do. It seems they have a belief with the same content. Both would certainly be able to straightforwardly answer the question, "what causes the tides?", although one could provide much more detail than the other. But what if they spoke different languages, would the content be the same then? Also, clearly it is not the exact same belief, but you might say that they are both tokens of the same type of belief.

Whether it is the same type of belief depends on the context. A dog and a human might both be attracted to the kitchen by the smell of frying sausages. Assuming that the dog's behaviour is driven by a state that plays the role of an 'intervening variable' (Whiten, 1996), then one might be tempted to say that both the dog and the human go to the kitchen because they believe that food is available there. Even if one is inclined to attribute a belief to the dog, could you maintain that the beliefs have the same content? They are likely stored in different formats and the human's belief is far more inferentially rich. But one could argue that in the context of attracting, or causing the behaviour of going to the kitchen, their belief is the same.

I can't hope to create an exhaustive notation for all possible ways in which the term could be used. More than three notational devices would begin to be overly confusing and would cause a cluttered text. To try to apply the notation too extensively would become restrictive, and would clash with ordinary practices, such as using italics for stress. To avoid these issues, I will use this notation sparingly, primarily where there is the possibility of confusion between the concept (BELIEF) and the word ('belief'). Where I explicitly refer to 'the concept of belief' I am clearly referring to the concept and will not use the BELIEF notation, although sometimes I will use BELIEF where spelling out 'the concept of belief' would make the sentence unwieldy. I will also occasionally use the notation where I need to distinguish the state or whatever it is that constitutes belief (*belief*). Beyond this, context and, if necessary, explanatory notes should suffice to make the meaning clear.

1. Eliminativism: The Case Against Belief

While many pre-scientific concepts used to understand the world have been rendered obsolete by advances in science, the concept of belief has, alongside the other key folk concepts of mind, continued to play a key role in theorising about the mind up to the present day[4].

But is the concept of belief's time finally up? Do recent advances in cognitive science show that folk talk about belief is obsolete? Should it, at least in scientific theorising, be eliminated and replaced with new concepts?

One theorist that argues for this position is Jenson (2016). In his argument in favour of elimination he appeals to Wimsatt's (1982) concepts of *robustness* and *fragility*, which holds a theoretical entity to be robust if it can be consistently demonstrated using a variety of different means of detection. If different means of detection give different results, then that theoretical entity is held to be fragile. Jenson argues that theoretical entities that can be shown to be sufficiently fragile do not exist and should be eliminated. He then reviews several studies that show a stark contrast between beliefs that are self-reported and beliefs as revealed by non-verbal behaviours. Jenson takes this to show that belief is fragile and therefore suitable for elimination.

In this section I will set out Jenson's claim and show that, although his use of a robustness analysis is inappropriate, there is a genuine threat to the use of the concept of belief in cognitive science.

In section **1.1** I identify Jenson's starting assumptions and discuss the relationship between folk psychology and cognitive science. In **1.2** I sketch Jenson's challenge to belief showing how it is different from the radical eliminativism of the 1990s. In **1.3** I introduce the concepts of robustness and fragility, and consider how and why we would eliminate certain concepts. In **1.4** I set out Jenson's target account of a belief. In **1.5** I review the evidence that Jenson gives to demonstrate that the concept of belief is fragile. In **1.6** I demonstrate that a robustness analysis is unable to demonstrate that the concept of belief should be eliminated. In **1.7** I describe a case study of the debate around the use of the concept of innateness in the cognitive sciences. I then draw several lessons from this debate to apply to the current inquiry. Finally, in section **1.8** I show that there is still a threat to the use of the concept of belief in cognitive science. I introduce an alternative framework within which I evaluate whether the concept of belief has a role to play in cognitive science.

1.1 Starting assumptions – folk psychology and cognitive science

There are two broad approaches that cognitive science can take to folk psychology, one conservative the other radical. The conservative approach would be to initially adopt the categories of folk psychology and to adapt and refine them as our knowledge progresses. The radical approach would be to sweep away folk psychological concepts and begin theorising from scratch.

[4] For example, the concept of belief is key in the recent literature investigating whether preverbal infants can understand false beliefs e.g. Onishi and Baillargeon (2005), Carruthers (2013a), and Southgate and Vernetti (2014). This is something I will explore in later chapters in chapters 4 & 5

The radical approach – eliminative materialism

Husband and wife Paul and Patricia Churchland are the best-known advocates of the radical approach. Their theory of eliminative materialism (e.g., Churchland, 1981) holds that folk psychology is a radically false theory about the way that our minds work.

The familiar mental categories of folk psychology (belief, desire, intention, etc.) are the posits of this radically false theory and therefore they have no place in a future scientific theory of mind.

This kind of eliminative materialism I shall refer to as 'radical eliminative materialism'.

The conservative approach – integrationism

The conservative approach, which following Frankish (2004, p.1-2) I shall refer to as **integrationism**, is the idea that folk psychology is such a successful way of explaining and predicting human behaviour that it must be at least a reasonably accurate description of what is actually going on inside our heads.

The familiar folk psychological mental state concepts then, are likely to form a basic taxonomy of mental states in a respectable scientific theory of the mind and the basic cognitive architecture suggested by folk psychology is likely to be broadly that used in a mature cognitive science (see, for example, Fodor,(1987) for this kind of argument).

Of course, most integrationists are not so extreme as to consider folk psychology to be the final word on the science of the mind. Folk psychology is a folk theory and folk theories have had mixed success when facing scientific scrutiny.

However, as a science has to start from somewhere, I shall maintain that the most sensible approach would be a form of open-minded modest integrationism – to start with the cognitive architecture of folk psychology as a working model and to continually revise it in light of evidence from scientific investigation.

This is likely to mean that as science progresses, we are going to see a divergence between scientific theories of the mind and folk psychology. How far these two positions will diverge isn't yet clear. It could be that only minor refinements are needed, or it could be that more fundamental changes are required. We need to keep an open mind, and we shouldn't be surprised if such refinements involve altering the definition of certain concepts in counter-intuitive ways or replacing them altogether.

1.2 Jenson's challenge

In this thesis I intend to determine whether the concept of belief should be eliminated from cognitive science. What makes Jenson's challenge interesting is that it starts from the modest integrationist position. While radical eliminativists would readily agree that belief should be eliminated, for integrationists it is an open question. Jenson intends to persuade moderate integrationists that the science has reached a point at which the refinements of an integrated folk psychology tip into the elimination of belief.

Note that in the above paragraph I used the term belief without the notation I discussed in the introduction. That is because the belief that Jenson thinks should be eliminated is both the word 'belief' and the concept of belief. Eliminating the term 'belief' from cognitive science does not guarantee that the concept will also be eliminated. It is possible that the concept, or something very close to it, could continue to play a role in cognitive science under a new label. This is something that I will return to later in the thesis.

As discussed above, integrationism takes the basic cognitive architecture of folk psychology as a starting point and continually refines it as our scientific understanding advances. Jenson believes that this refinement has reached a point where we need to eliminate the concept altogether. This makes his position distinct from the radical eliminative materialism of the 80s and 90s. His is not an argument that folk psychology is a bad theory that will be replaced wholesale by a new, scientific theory of the mind. Rather he argues that recent scientific insights have shown that folk belief conflates several distinct states that should be treated separately in scientific discourse and that the umbrella term of 'belief' and the broad concept of belief should be abandoned.

The consequences of this type of elimination would not be the "greatest intellectual catastrophe in the history of our species" as Fodor (1987, p.*xi*) thought would be the case if our folk psychology turned out to be incorrect. Rather, in Jenson's eliminative scenario, the folk would likely continue much as before while scientists would replace belief with two or more new states that will increase the explanatory and predictive power of their theories.

Radical eliminative materialism has fallen from favour over recent years, but even those who reject it should still take Jenson's challenge seriously.

Note that Jenson argues that the concept of belief should be eliminated from cognitive science, but it doesn't follow that it should be eliminated from our everyday understanding of other minds. Against radical eliminative materialism, some argued that folk psychology isn't a theory, therefore its categories are not suitable candidates for elimination (e.g. Hannan, 1993; Wilkes, 1993). But Jenson sidesteps these arguments by drawing the distinction between the folk psychological concept of belief and the concept of belief as it operates in cognitive science (Jenson, 2016, p.2). We could concede for the sake of argument that folk psychology isn't a theory when it operates in our everyday interactions, but when the integrationist transplants folk psychological concepts and architecture into cognitive science, it is acting as part of a scientific theory that aims to describe how our minds actually function. In this context it is very much open to elimination as our knowledge advances.

As evidence that the concept of belief should be eliminated, Jenson sets out to demonstrate that it is **fragile**. That is to say, it cannot be detected by using multiple independent means of detection (Wimsatt, 1982).

As an exemplar of the folk psychological concept of belief, Jenson identifies Schwitzgebel's (2011) summary of the functional role of the concept of belief. This functional role suggests two ways in which the presence of an agent's belief can be detected – verbally and non-verbally. Jenson then details several studies from the scientific literature that show that the verbal and non-verbal methods of determining an agent's possession of a belief fail to agree. Jenson argues that this shows that the concept of belief is fragile and that it should therefore be eliminated from cognitive science.

In sections **1.3** to **1.5** below, I will set out Jenson's challenge in more detail. At this stage I will not offer any criticism, but in subsection **1.6** I will demonstrate that a robustness analysis fails to show that belief should be eliminated from cognitive science.

1.3 Robustness and fragility

As discussed in the previous section, Jenson argues that the concept of belief should be eliminated because it is fragile. To properly evaluate the argument, we need to introduce the concept of **fragility** and its opposite **robustness**.

Wimsatt (1981, p.126) gives the following procedure for a robustness analysis:

1) Analyze a *variety* of *independent* derivation, identification, or measurement processes
2) Look for and analyze things that are *invariant* over or *identical* in the conclusions or the results of these processes
3) Determine the scope of the processes across which they are invariant and the *conditions* on which their invariance depends

4) Analyze and explain any relevant *failures of invariance*."

To adapt a simplified example from Jenson (2016): I'm wandering the desert and I see a body of water on the horizon. Experienced desert wanderers will know not to get too excited at this point, as I am detecting the water using a single method of detection – vision. What I see could indeed be a body of water, an opportunity to slake my thirst. However, it could also be an illusion, an artefact of the single mode of detection that I am using. It is well known that high temperatures and flat surfaces can cause reflections that give the illusion of water.

It is only when I am able to approach the apparent pool and use other means of detection (by touching and tasting it) that I can confirm the reality of the pool.

Robustness analysis has had its critics. For example, Woodward (2006) identifies several different types of robustness analysis and takes aim at what he identifies as 'inferential robustness'. Jenson argues that the variety of robustness that we are discussing here is the less controversial 'measurement robustness', although he frames measurement as 'detection'.

For Jenson's measurement robustness analysis, we start with a target entity of some kind. If we are able to reliably detect evidence of this target entity using a variety of different detection methods then our concept or category for this entity has been shown to be **robust** and we can be more confident that what we are detecting has some basis in reality. However, if across our different means of detection, some are able to detect our target entity while others can't and we are unable to explain these variances, then the our category has proven to be fragile. We should have less confidence that fragile categories have some basis in reality.

The reason for the intuitive appeal of measurement robustness analyses is that while each different detection method may have its own type of error, very different methods of detection are unlikely to have exactly the same error. If very different detection methods agree on a measurement, then it makes it highly likely that the common measurement is robust i.e., it is accurate. Thinking back to the desert example, while it is possible that my eyes may be deceived into detecting water by the mirage, it is highly unlikely that there will also be a co-occurrence of other phenomena that deceives my other senses into thinking that water is present.

The example cited by both Jenson (2016) and Woodward (2006) as an example of a successful robustness analysis is Perrin's (1916) work to support the reality of atoms. Perrin worked to establish the reality of atoms by establishing the number of atoms in a mole of a substance – Avogadro's number. He invented several different methods to measure Avogadro's number, and to these methods several were added by others. "Perrin counts thirteen different experimental techniques including those with a basis in Brownian movement, alpha decay, X-ray diffraction, blackbody radiation, and electrochemistry" (Jenson 2016, p.5). All of these measurement techniques produced a very similar figure. As discussed above, while each of these techniques may have their own peculiar error, it is unlikely that such diverse methods would have the same error. Avogadro's number was therefore established and the reality of atoms supported[5].

Jenson flips the robustness analysis on its head. His aim is not to establish robustness to support the existence of a something, but to establish the absence of robustness – **fragility**, to establish that the putative existence of something is not adequately supported by the evidence. If something is not supported by the evidence, then it may well be a candidate for elimination.

[5] The exact form of Perrin's reasoning and its implications for the case for realism of atoms remains the subject of debate, see e.g. Chalmers (2011) and Hudson (2020). I won't be drawn into this debate as I will adopt a different strategy to show the failure of Jenson's robustness analysis.

It is worth noting immediately that while establishing robustness gives us strong support for the existence of something, establishing fragility does not give the same kind of strong positive support in favour of elimination. Robustness analysis gains its appeal from the fact that diverse methods of detection are unlikely to have the same error. If diverse methods of detection do not agree, this could be due to the entity that we are trying to detect not existing or it could also be due to the errors in each detection method. Jenson notes (2016, p.970) that:

> "If Perrin had gotten variant results across different experimental techniques, it would have been rational for his contemporaries to be much less confident in the existence of molecules. Furthermore, given that the existence of molecules was controversial at the time, it would have been reasonable to think that it is more likely that they do not exist than it would be to suppose that something had gone wrong with the experiments."

Variant results clearly fall short of establishing that atoms do not exist, but it should rightly raise doubts about the existence of the entity under consideration. As an example of elimination by fragility, Jenson (p972) makes use of Kuhn's (1962) classic example of phlogistic theory. Phlogistic theory made several predictions that weren't observed in the data. For example, when a substance burned it was supposed to lose phlogiston into the air, the theory would therefore predict that burned substances should lose mass, but the opposite was true in the case of burned metals. Similarly, the theory predicted that the mass of air would increase when a substance was burned. This was the case when many materials burned, but again the opposite was true if metals were burned.

1.4 What is a belief?

In order to perform a robustness analysis Jenson needs to give an account of belief. The most common way of marking out beliefs from other mental states is by their functional role. While noting that there has been little work done to precisely define what the functional role of belief is, Jenson (2016, p972) singles out Schwitzgebel's summary as an exemplar of what analytic philosophers broadly take to be the functional role of belief (from Schwitzgebel, 2011):

> 1) Reflection on propositions (e.g., [Q] and [if Q then P]) from which P straightforwardly follows, if one believes those propositions, typically causes the belief that P
>
> 2) Directing perceptual attention to the perceptible properties of things, events, or states of affairs, in conditions favorable to accurate perception, typically causes the belief that those things, events, or states of affairs have those properties (e.g., visually attending to a red shirt in good viewing conditions will typically cause the belief that the shirt is red)
>
> 3) Believing that performing action A would lead to event or state of affairs E, conjoined with a desire for E and no overriding contrary desire, will typically cause an intention to do A
>
> 4) Believing that P, in conditions favoring sincere expression of that belief, will typically lead to an assertion of P.

Points 3) and 4) of this description suggest two different methods for detecting an agent's belief – 3) suggests that the presence of beliefs can be detected by observing an agent's non-verbal behaviour, and 4) suggests that beliefs can be detected by simply asking the agent what they believe. If these two methods of detecting belief can be shown to come up with different results, Jenson argues that the concept of belief (at least as it is described in Schwitzgebel's summary) will be shown to be fragile and a potential candidate for elimination. In the next section, I take a look at some of the studies discussed by Jenson that show exactly that.

1.5 Evidence of fragility?

In the last section I identified two different methods of detecting an agent's beliefs, firstly by looking at their non-verbal behaviour, and secondly by asking them what their beliefs are. In this section I review three studies described by Jenson (confabulation, contamination, and implicit association tests) that show that these two detection methods can come up with different results. I also add a fourth study (a type of false-belief task) that also demonstrates that different detection methods can have different results.

Confabulation of reasons

In a famous study by Nisbett and Wilson (1977, p.243-244) participants were presented with four identical pairs of nylon stockings and asked which of the stockings they preferred. Participants tended to select the stockings that were the furthest to the right. As the stockings were identical, the results suggest that people have a right-side bias. When asked why they chose the particular pair of stockings they did, no participant mentioned the positioning of that stocking relative to the others. When the experimenter asked about whether the positioning of the stockings had an effect on their choice, virtually all of the participants denied that it had. Jenson argues that the non-verbal behaviour of arbitrarily selecting the stockings on the right suggests that at some level they believe that the stockings are identical. However, the participants' verbal responses, confabulating reasons for their choice and denying that positioning had any influence, suggest that they also genuinely believe the stockings were different. Here we have two different detection methods coming up with different results – support for the case that the concept of belief is fragile.

Contamination

The work on contamination by Paul Rozin and colleagues provides further examples where non-verbal behaviour suggests that an agent holds a particular belief, but the agent explicitly states that they hold no such belief. In one such study (Nemeroff and Rozin, 1994), participants were asked to eat fudge shaped like dog faeces and drink lemonade from a bedpan. Participants were reluctant to do this, despite the fact that they knew that it was fudge and that the bedpan was clean. Jenson again argues that this shows that at some level, participants believe that the fudge is faeces and that the lemonade is urine. Yet, when asked, the participants denied that they had such a belief. Again, two different methods of detecting an agent's belief come up with different results.

Implicit association tests

The final type of evidence discussed by Jenson comes from implicit association tests (IATs) (e.g., Greenwald *et al,* 2000). In the first part of the IAT, participants are asked about their attitudes to certain sections of society, for example dark-skinned or light-skinned people[6]. In the second part, participants take part in a computer-based task in which they are shown a series of faces and have to sort the faces by clicking one of two keys. They are then shown a series of positive and negative words and have to categorise them again using the two keys. Next, faces and words are mixed together and participants need to categorise them using just two keys, for example the 'E' key for positive words and light-skinned faces and the 'I' key for negative words and dark-skinned faces. The test is repeated several times with the keys and the categories mixed up in all possible combinations.

[6] You can take a test here – https://implicit.harvard.edu/implicit/takeatest.html

In a meta-analysis of IATs, Greenwald *et al* (2009) found that the explicit statements of belief did not accurately predict the results of the computer-based association task. This means that many participants who stated that they don't prefer one group over the other, were shown in the test to have a definite preference for one group. According to Jenson, the results suggest that despite subjects' explicit statement that they are egalitarian, at some level they hold racist beliefs.

Once again, agents' stated beliefs differ from those revealed by their non-verbal behaviour.

False belief tasks

I will add one slightly different example to Jenson's list of studies that I think helps to show even more clearly that different detection methods yield variant results. Clements and Perner (1994) devised a variation of the traditional children's false belief task first seen in Wimmer and Perner (1983). The aim of false belief tasks is to determine at what age children gain an understanding of the concept of belief. If children are able to demonstrate an understanding that an agent can have beliefs that differ from reality (and their own beliefs), then they are taken to have the concept of belief.

The way that this is assessed in the original false belief task was to have a puppet place an object in location A and then leave the scene. While the puppet is out of the scene, the object is moved from location A to location B. When the puppet returns to the scene to retrieve the object, test subjects are asked to verbally state where the puppet will first search for the object.

The correct answer is clearly location A. While the object has been moved to location B, this occurred out of sight of the puppet and so the puppet has the false belief that the object is in location A. Children only start to pass this test at around four years of age.

In the Clements and Perner variation on the false belief task, children aged between two years five months and four years six months were shown a short scenario involving puppets and a block of cheese. The set-up is similar to the classic test, a puppet leaves a block of cheese in location A and, in the false belief condition, it is moved to location B while the puppet is unable to observe the switch.

The difference between the Clements and Perner version of the task and the original version is that when the puppet returns to the scene and the test subject is asked where it will look for the cheese, the gaze direction of the test subject was recorded. Somewhat surprisingly, of the children aged over two years 11 months, 90% looked in the correct location (location A, where the puppet incorrectly thought the cheese was) while just 45% verbally responded with the correct location.

For many of these test subjects, their gaze behaviour suggests that they have the belief that the puppet will look in location A, but their explicit answer suggests that they have the belief that the puppet will look in location B.

This task is another example of an agent's verbal behaviour suggesting one belief and their non-verbal behaviour suggesting another. Different modes of detection yield different results.

What is different about this task is that while we could argue in the other examples that the non-verbal behaviour is guided by an aversion, a preference or a bias, in this example the child is demonstrating an expectation of where the puppet will look. Additionally, the behaviour is in response to a direct question from the tester. This shows more clearly than the other examples that the test subjects have what appear to be conflicting beliefs. I will return to this later in the thesis.

Having shown in several empirical studies that different detection methods yield different results, Jenson concludes that the concept of belief is shown to be fragile. In the next section I suggest that this conclusion is not justified.

1.6 Is belief fragile?

In this section I aim to show that a robustness analysis is an inappropriate framework in which to judge whether belief should be eliminated. I will show that belief is a flexible concept, and that the properties of the functional role set out by Jenson does not constitute the list of necessary conditions which it needs to be for the robustness analysis to succeed.

This will lead me in section **1.8** to develop an alternative framework to the robustness analysis in which to consider the question of whether belief should be eliminated.

While I accept that the empirical studies show that different detection methods yield different results, I argue that the concept of belief, as set out in Schwitzgebel's functional definition, is flexible enough to tolerate a diversity of types or species of belief. Rather than showing that belief is fragile, the empirical evidence from the studies discussed could simply be detecting beliefs that do not have all four of the properties listed.

Consider again the list of properties of belief that Jenson takes as indicative of the functional role of belief (Schwitzgebel 2011):

1) Reflection on propositions (e.g., [Q] and [if Q then P]) from which P straightforwardly follows, if one believes those propositions, typically causes the belief that P

2) Directing perceptual attention to the perceptible properties of things, events, or states of affairs, in conditions favorable to accurate perception, typically causes the belief that those things, events, or states of affairs have those properties (e.g., visually attending to a red shirt in good viewing conditions will typically cause the belief that the shirt is red)

3) Believing that performing action A would lead to event or state of affairs E, conjoined with a desire for E and no overriding contrary desire, will typically cause an intention to do A

4) Believing that P, in conditions favoring sincere expression of that belief, will typically lead to an assertion of P.

Take property 3) as an example. Firstly, for a belief to result in an action, a specific set of conditions needs to obtain. The belief that performing action A would lead to event or state of affairs E needs to be paired with a desire for E. There also needs to be no desire that overrides E. To further complicate matters, we need to consider the conditions that need to obtain for the intention to do A to result in performance of the action A. The description here is of a contingent connection between belief and action – possession of a belief does not necessarily result in the formation of an intention to act.

Secondly, even where all of the correct conditions obtain, according to Schwitzgebel's definition, such a belief only *typically* causes an intention to act. The same is true of all of the functional properties including 4) – believing that P, where the correct conditions obtain, only *typically* leads to the assertion of P.

The use of the word 'typically' implies that the functional role being described is that of a typical or paradigmatic belief. It is certainly not a list of necessary conditions. Therefore, from the fact that a state lacks the functional properties 3) or 4) it doesn't follow that that state is not a belief, it only means that it cannot be a typical belief, i.e., it does not exhibit all the properties on Schwitzgebel's list.

The description of belief's functional role clearly has some flexibility and leaves room for some variety in the type of states that can be considered to be beliefs.

Consider the variance in the implicit association tests. The first method of detection is the subjects' verbal behaviour – test subjects affirm that they hold no racist beliefs. Yet, on the second method of detection their non-verbal behaviour as revealed in the test shows that, arguably, they hold racist beliefs. To explain this variance, one could argue that the state that is behind the non-verbal behaviour revealed in the IAT is a non-typical belief. The state could have properties 1), 2) and 3) but lack property 4), i.e., it does not lead, even where the correct circumstances obtain, to a verbal assertion that P.

We have established that the functional role of a belief as set out in Schwitzgebel (2011) and used by Jenson to identify detection methods for his robustness analysis is flexible, and does not rule out the possibility of non-typical beliefs that do not have all four of the listed functional properties. For this reason, the evidence reviewed does not conclusively prove that belief is a fragile concept.

Indeed, if there are such things as non-typical beliefs that don't exhibit all of the functional properties listed above, then it is likely that they would be revealed by just the type of non-typical experimental circumstances that Jenson discusses in his evidence.

The robustness analysis fails to show that belief should be eliminated because the functional definition used by Jenson is tolerant of a diversity of states within the concept of belief. The variances in detection highlighted by his empirical examples do not show that belief is a fragile concept, they showcase its diversity.

However, while Jenson's robustness analysis fails to show that the folk concept of belief is fragile, it does not follow that the folk concept is suitable to be used unchanged in cognitive science. Perhaps the concept of belief has to be so flexible because it contains distinct categories – Distinct categories that perhaps scientists would be better served by treating separately.

What we need is an alternative framework within which we can properly evaluate the eliminativist challenge to belief. Before we move on to this, it will be instructive to briefly move away from talk of belief to consider a similar debate within the philosophy of cognitive science. In **1.7**, I will examine the debate around the elimination of the concept of **innateness** in cognitive science.

1.7 A case study – innateness in cognitive science

It will be useful to consider an example of elimination from a different field as a case study. In this section I will consider the debate surrounding the concept of innateness and its use in cognitive science. I will set out the debate before identifying several lessons that we can learn from this parallel discussion. I will then apply these lessons in constructing an alternative framework to the robustness analysis in **1.8**.

While the concept of innateness is not quite as embedded in our everyday lives as the concept of belief, the idea that some traits of organisms are part of the essential nature of an organism while others are learned or acquired is part of our shared cross-cultural folk biological knowledge (Mameli & Bateson, 2006).

As well as being part of a biological folk theory, innateness has a long history of use in scientific discourse. The concept has been supplanted in many areas of biology although it remains an important concept in the field of immunology[7]. The concept of innateness also has a recent history of use within cognitive science and its elimination from this specific domain has been the subject of recent debate.

[7] The innate immune response is contrasted with the adaptive immune response. Roughly, the adaptive immune response is pathogen specific and develops after exposure to different pathogens, whereas the innate immune response does not change in response to pathogen exposure (Roitt et al 2001)

Like the debate about the concept of belief in cognitive science, the debate about innateness in cognitive science turns on whether the folk concept can be sharpened up sufficiently for it to function as a useful concept within cognitive science.

The case for eliminating the concept of innateness

Using experimental philosophical techniques, Griffiths et al (2009) established that what determined whether subjects thought that a trait was innate was influenced by at least two broad features:

1) **Developmental fixity –** the development of the trait is insensitive to the environment

2) **Species typicality –** every normal individual of a particular type of organism (or a specific subcategory of that type of organism e.g. male, adult etc.) has the trait[8].

Mameli and Bateson (2006) argue that the challenge for those who want to preserve the concept of innateness in cognitive science is to develop a scientific concept that captures the essence of the folk concept. They run through 26 potential scientific successors to the folk concept. All are open to counterexamples and none of them are satisfactory on their own as a potential scientific definition of 'innate'.

They attempt several precise definitions for developmental fixity, including "A trait is innate if and only if its development doesn't involve the extraction of information from the environment", which fails because of the lack of a principled distinction between "environment-as-information vs environment-as-support" (Mameli & Bateson, 2006, p.161).

Species typicality also fails as a definition. Certain genetic diseases are not species typical, but we would say that they were innate. Conversely, there are species typical traits that we would not consider innate. For example, several bird songs are species typical, but are learnt during young age from conspecifics.

These two examples also show that the two broad features of the folk concept of innateness are dissociable from each other. The genetic disease is not species typical but it is developmentally fixed, while the bird song is species typical but is not developmentally fixed.

The folk concept of innateness then, conflates several independent and dissociable properties. The major negative effect of this conflation is that it encourages unwarranted inferences. If a trait is labelled as innate, then we could infer that it possesses all the properties of innateness. But as we have seen, the properties are dissociable. We could also be tempted to infer from the fact that a trait has one of the features of innateness, say species typicality, that it also has the other properties associated with innateness. But again, the examples discussed above show that this would be unwarranted.

Due to the confusion and potential for unwarranted inferences, the folk concept of innateness should have no place in cognitive science. It should be replaced by several, more specific terms, that do not have the gamut of connotations that innateness has.

Attempts to rescue the concept of innateness

Could we save the concept of innateness by identifying it with one of these successor concepts? There have been attempts to do just this. Griffiths et al (2009) discuss the efforts of Ariew (1996) to reform the concept of innateness and identify it with the biological concept of 'environmental canalisation' – essentially the extent to which a trait will develop despite variation in the environment.

[8] They found that a third feature that they had hypothesised was part of the folk concept of innateness, 'purposive function', was less important to whether subjects were willing to apply the innate label to a trait.

However, scientists already have the concept of environmental canalisation, what additional benefit to them is identifying it with the old concept of innateness? In fact, this identification has the potential to mislead. If scientists were to communicate their work to the public making use of the new refined concept of innateness, identical in definition to environmental canalisation, then there is a high risk that non-biologists will make the kind of unwarranted inferences that we discuss above.

Additionally, the folk concept is clearly far richer than the concept of environmental canalisation. The difference between the two concepts is so large that it makes it hard to argue that the concept of environmental canalisation is a refinement of innateness, rather than an attempt motivated by saving the term 'innateness'. How far a concept can be refined and still be the same concept is discussed in Mameli and Bateson (2006) who discuss the amount of "overlap" that can be allowed between a folk concept and its scientific successor for them to be considered to be genuinely picking out the same natural phenomenon – "the bigger the mismatch, the more problematic is the alleged theoretical reduction." (Mameli & Bateson, 2006, p.157).

Another attempt to save the concept of innateness is made by Richard Samuels (2007). Samuels emphasises the distinction between the **word** innate and the **concept** innate. The word, Samuels agrees, is ambiguous. It is used in a variety of different ways, there is the folk usage and then there are different usages across several scientific disciplines. But it doesn't follow from this that the concept of innateness is confused.

Samuels also makes it clear that he is aiming to explicate a concept of innateness that is used in cognitive science, rather than as a general scientific term to be useful across all scientific disciplines.

The thrust of Samuels' argument is that the properties that various theorists have argued are conflated in the innateness concept (what Samuels calls I-properties)[9] are not constitutive of the concept of innateness but simply provide evidence that some trait may be innate. "Roughly put, the occurrence of I-properties is not a (conceptually or metaphysically) necessary condition for something's being innate. Rather, it is merely that discovering a trait possesses one or more I-property provides *evidence* that the trait is innate." (Samuels, 2007, p.23).

One way in which this evidential relationship could hold is if innateness were a natural kind. Samuels draws on the work of Richard Boyd, who characterises natural kinds as homeostatic property clusters (Boyd, 1991).

According to this idea, a natural kind is (from Samuels, 2007, p.23):

1) Associated with a range of characteristics or symptoms which tend to be co-instantiated by instances of the kind, but are not genuine necessary conditions for membership
2) There is some set of underlying causal mechanisms and constraints – "causal essence," if you will – whose operation explains the co-instantiation of these various symptoms
3) To the extent that there is any real definition of what it is for something to be a member of the kind, it is not symptoms but causal essence that defines membership.

Samuels illustrates the idea of natural kinds as a homeostatic property cluster with the example of influenza. While having the flu involves presenting a range of symptoms (coughing, temperature, sore throat etc.) presenting those symptoms count as evidence that you may have the flu, but they do not

[9] Samuels list five I-properties but for our purposes we can make do with the two major features of innateness identified by Griffiths et al 2009

constitute a list of necessary conditions for having the flu. One could still have the flu if one didn't have a sore throat, for example.

The causal mechanism underlying the various symptoms is infection by the influenza virus. And it is this infection that determines whether someone has the flu.

Samuels goes on to identify innateness with psychological primitiveness, and the range of I-properties as the equivalent of symptoms providing evidence of innateness, rather than constituting it.

Samuels develops a convincing argument concerning the difference between the evidence for natural kind membership, necessary and sufficient conditions, and also that what theorists in cognitive scientists are interested in is often what is psychological primitiveness. But does he rescue the concept of innateness? He aims to modify the concept to make it scientifically respectable, but his solution is so far removed from the folk concept of innateness that it is arguably a different concept.

His solution also seems susceptible to the point made by Griffiths et al (2009) discussed above, if you are identifying innateness with the existing concept of psychological primitiveness, why not just call it psychological primitiveness? The term 'innate' has so many connotations and baggage, and the risk of unwarranted inferences is so great that it seems sensible to abandon the label altogether for a scientifically respectable successor term. While Samuels' argument is convincing, his preservation of the 'innate' label is less so.

Lessons to be applied to the belief debate

There are several lessons that we can learn from the literature on the concept of innateness that we can apply to our present inquiry into the elimination of belief.

1) **Elimination can be domain specific** – Debates around the integration or elimination of folk concepts need not be a simple dichotomy between folk concept and scientific concept. Innateness has been eliminated across most biological domains although it remains a key concept in immunology and has usage – albeit controversial – in cognitive science.
2) **Properties can be evidential rather than constitutive** – Samuels' argument shows that properties that seem to constitute a folk concept can be construed as evidential rather than constitutive of a refined scientific concept. This could be the case for example, if the concept were a natural kind.
3) **Saving a label or term from elimination is not the same as saving a concept from elimination** – In developing a potential scientific successor concept to a folk concept, we should be aware that the further away that refined concept is from the original folk concept, the more of a stretch it is for it to be considered the same concept. For refined concepts that are so far removed from their folk origin, using the same term can cause more confusion than good. This is especially the case where a folk concept is identified with an existing scientific concept that already has a name and a history of usage within scientific discourse.

1.8 A new framework for elimination

As demonstrated in subsection **1.6**, Jenson's robustness analysis fails to show that belief should be eliminated. However, there is still a case for the elimination of belief that needs to be answered. While the experimental data discussed by Jenson does not show that belief is a fragile concept in the way that he wants it to, it does show that there may be several different states within the category of belief that have quite different functional roles. Indeed, the reason that the robustness analysis fails is that the functional role of the folk concept of belief is tolerant of a diverse range of non-typical beliefs.

Diversity within general terms is to be expected and it can be tolerated without thinking that we need to eliminate the general term. Take the general term 'sculpture', a variety of different objects can fall under this heading. Sculptures can be figurative or abstract, they can be of any material or combination of materials. While the English word sculpture derives from the Latin 'sculpere' to 'carve', sculptures can also be cast, modelled or assembled from pre-existing parts. The key aspects of what make something a sculpture are that it has been created by a human hand, that it is in some way three dimensional and that it was created for aesthetic or artistic purposes. Something designed to be functional, a building for example, cannot be sculpture, but can feature sculptural elements – elements that have non-functional decorative forms such as the ornate capital of a Corinthian column or a grotesque on a gothic cathedral.

It is not just everyday or folk terms that are tolerant of diversity, many scientifically acceptable terms also have a diverse membership. The category of liquids, for example, contains liquid nitrogen, mercury and my cup of coffee. While in many respects these are very different substances, it is their similarity in certain respects, namely that they are near non-compressible fluids, that unites them under a single category.

That there is a diversity of states that fall under the folk heading of belief is therefore not sufficient reason for the category to be eliminated. Even if there were several clear groups of states within the category of belief, this again is no sufficient reason to eliminate belief. Science is comfortable with hierarchies of categories, with categories containing subcategories that themselves contain subcategories. Taxonomy in biology is a notable example of this.

The real challenge then, is the claim that the folk category of belief conflates several distinct belief-like states and the general term becomes redundant. A broad term that conflates distinct states may be fine for everyday usage, the folk concept of belief has been guiding human behaviour for several thousand years at least. However, what is fine for everyday usage may not be appropriate in a scientific context. The sharpening up process that we described earlier may require us to distinguish the conflated belief-like states and treat them separately in our theorising. If this is the case, then what role should belief play in cognitive science? There are two options: we could maintain belief as an umbrella term for the now distinct belief-like states or we could eliminate the concept from science altogether.

In this subsection, taking into account the lessons learned from the innateness debate, I will set out a new framework within which to evaluate this challenge to belief. I propose that whether a category should be eliminated or not should be determined by applying what I call the utility test. I will then distinguish two different types of elimination – pragmatic elimination and ontological elimination – and describe how we apply the utility test to determine which, if either, type of elimination is appropriate.

How then should we determine whether a concept or category should be maintained or whether it should be eliminated? I propose a simple test:

The utility test - Is it useful to distinguish things of kind K from things that are not of kind K?

If there is no use in distinguishing a category, then it should be eliminated. The utility test is domain relative. While it might be useful to distinguish things of kind K in one context, it may have no use or even be downright misleading in others. Alternatively, it may become apparent that there is no domain in which it is useful to distinguish things of kind K, in this case it is appropriate to eliminate the category completely.

The domain relative nature of the utility test suggests two different types of elimination, what I will call **pragmatic elimination** and **ontological elimination**.

Pragmatic elimination – In cases where a category has no use in one domain yet retains uses in other domains, we can say that the category is pragmatically eliminated from the domain in which it has no use. The category may continue to be used unaltered in domains where it does have value.

Ontological elimination – In cases where there is no use in distinguishing a category in *any* non-fictional domain, then we can say that this category has been ontologically eliminated. It simply does not exist.

As an example of pragmatic elimination consider the category of nuts. If I am coordinating snacks for a gathering of people and I ask someone to bring a selection of nuts, I am setting a reasonably clear expectation of what they should bring. A standard selection would include favourites such peanuts, cashews and pistachios. Perhaps if they were feeling especially lavish, we could also expect almonds or walnuts. We have a shared folk conception of the category of nuts[10] formed around their properties as foodstuffs. They are dry, oily, have a distinctive type of crunch and, despite a great diversity of flavour, share what we would vaguely call a 'nutty' flavour.

From a botanical point of view however, classification on the basis of superficial similarity in look and taste is not sufficient, a more technical definition based on specific properties and evolutionary lineage is necessary. If I describe a new species of plant as producing nuts then my fellow botanists will be able to infer that the tree produces dry fruits that do not split open to reveal their seed at maturity (https://www.britannica.com/science/nut-plant-reproductive-body accessed 12/2/2020)

On the botanical definition of nuts, none of the food items brought to the party mentioned above are nuts. Peanuts are a legume, for example, while a pistachio is a drupe.

What we have here is two separate categories, although confusingly they both have the same label. The two categories can coexist with minimal confusion because it is usually clear which of the two categories is being referred to, based on the context and the community within which the term is being used. If there is any danger of confusion, for example if someone were asked to bring nuts to a party at the botany department of a university, then we could clarify by specifying 'culinary nuts' or 'botanical nuts'.

Despite the botanical definition of nuts, the folk concept of nuts still passes the utility test in certain social and culinary contexts. There is clearly still value in distinguishing between nuts and non-nuts. But in scientific discourse, the definition has been superseded. That is to say, the folk concept of nuts has been pragmatically eliminated from botanical discourse.

If there was no non-fictional context in which the folk category of nuts passed the utility test, then we would have a case of **ontological elimination**. This type of elimination is associated with the radical eliminativism championed by the Churchlands. For a concept to be ontologically eliminated, the category in question must be a posit of a radically false theory. Usually, the way that we know that a theory is radically false is because it has been replaced by a new theory, which has greater explanatory and predictive benefits. Returning to the example of phlogistic theory, discussed in Kuhn (1962) and by Churchland (1981), which posited a substance, phlogiston, that explained several phenomena observed by 18th century chemists.

Kuhn describes how phlogistic theory persisted despite it becoming increasingly difficult to explain new scientific observations in its terms. It was only after the development of the oxygen theory of combustion that scientists finally abandoned phlogistic theory. The new theory was so radically different to the old, discredited theory that there was no way that phlogiston could be salvaged. It couldn't be transferred into the new theory, no matter how much it was sharpened up. Phlogiston was ontologically eliminated, it simply did not exist.

[10] At least this is the case in English. Having had discussions with Spanish speakers I am told that there is no equivalent catch-all term for what I would call nuts. As well as being purpose relative, the utility test is clearly community relative.

Back to belief, clearly it passes the factual test. We refer to 'belief' all the time in non-fictional contexts[11]. As for the utility test, for everyday purposes reference to belief has clear value – it provides us with a quick and easy way of explaining and predicting the behaviour of friends and strangers alike. It helps us make sense of the social world that we inhabit. However, as I have established, while a term might have value for everyday purposes, it may have limited value or be downright misleading in a scientific context.

In this chapter I have argued that the concept of belief and the other concepts of folk psychology should, at least initially, be integrated into cognitive science. From this point, they should be constantly sharpened up to improve their predictive and explanatory capacities. This presents a threat to the concept of belief. Perhaps, as part of this sharpening up process, belief will be fractured into two or more successor states.

I evaluated the robustness/fragility framework proposed by Jenson (2016), finding that it is unsuitable to determine whether a flexible concept such as belief should be eliminated or not. While the evaluation framework was found wanting, I did determine that the concept of belief has a case to answer against the eliminativists. Using innateness as a case study, I showed how concepts that conflate several properties can be misleading and can seem to warrant misleading inferences. Based on lessons learned from the case study of innateness, I developed a new framework which is more suited to determine whether the concept of belief should be eliminated or not. That framework was based on the utility test – is it useful to distinguish things of kind K from things that are not of kind K? If it is not useful to distinguish a category or concept within a scientific discipline, then that category or concept should be pragmatically eliminated from that discipline. The key question of this paper then can be expressed as follows – does the concept of belief have scientific utility?

We have seen how belief is a diverse category. The major threat of elimination for belief is that the concept fractures into distinct successor states and the general term falls from use and becomes redundant. However, if belief can be shown to have utility in cognitive science it should resist elimination. Such utility may be only temporary. The general concept could act as a placeholder while a consensus about successor states emerges, or it could have a longer-term usage if no such consensus emerges. Even if a consensus about how belief should fracture into successor states does emerge, a general term that covers all of these successor states could still have a use in certain disciplines or debates within cognitive science, even if it is pragmatically eliminated from others.

[11] While we don't use the term 'belief' that frequently, we do have many belief-like terms. See chapter **2** for a discussion.

2. Folk ‘belief’ and philosophers’ BELIEF

In the previous chapter I set out a recent challenge that the concept of belief should be eliminated from scientific discourse. The challenge was unconvincing, however, because of the flexibility of the concept. I then set out an alternative framework within which the eliminativist challenge to the concept can be more effectively evaluated. To apply this framework and fully evaluate the eliminativist challenge to belief, it will be necessary to examine the concept in some detail to determine precisely what it is we are evaluating. This will be my task in this chapter.

In chapter **1** I set out two options for the relationship between cognitive science and folk psychological concepts such as belief – radical eliminativism and integrationism. Noting the longstanding and continuing success of folk psychological inferences in our everyday lives, I suggested that the most sensible approach for cognitive science would be to transfer the concepts of folk psychology, including belief, into scientific discourse.

In the example of innateness that I discussed in the previous chapter, philosophers were attempting to identify the folk concept of innateness with already existing scientific concepts – environmental canalisation for example. In cognitive science, there are, as yet, few commonly accepted successor concepts to folk psychological concepts and so integrationism provides a useful starting point for scientific theorising. As our knowledge of cognitive science improves, scientists could then, if necessary, sharpen up these concepts to maximise their explanatory and predictive capabilities. One element of this sharpening up process could involve the fracturing of general concepts into two or more distinct successor concepts, rendering the more general concept redundant and open to elimination.

But why would we think that belief and other pre-scientific concepts would be suitable to transfer into science in the first place? As Robert Cummins says “there is no evident reason why a serious empirical psychologist should care what the ordinary concept of belief is any more than a serious physicist should care what the ordinary concept of force is.” (Cummins, 2010, p.14). We should expect scientists to form their own concepts for their own purposes and these purposes are clearly very different from the purposes of the folk when discussing their daily lives and the lives of others.

Scientists often preserve the labels of pre-scientific concepts and attach them to scientific concepts. Examples of this include such words as ‘force’, ‘mass’ and ‘gravity’. But physicists’ concepts of force, mass and gravity are not shaped by what ordinary folk have to say about them. Additionally, pre-scientific concepts are likely to be too crude and multifaceted to be effectively delineated. In these cases, the terms are preserved but the concepts are not. So why would cognitive scientists care about the folk concept of belief and what philosophers have to say about it?

Against this presumption I will argue that the concept of belief is not a raw pre-scientific concept that is readily delineated from how folk use the word ‘belief’. Rather, as I will demonstrate in this chapter, the concept of belief is **doubly theoretical**. Firstly, the concept is a generalised abstraction from everyday usage of the term. Secondly, the concept plays a key functional role in the theoretically structured body of knowledge that underlies our mental understanding abilities.

The term ‘belief’ has been repurposed by philosophers to stand for the generalised **concept of belief** – a doubly theoretical general concept that subsumes the functions of a wide variety of cognitive states.

This status makes it a much more suitable to be integrated into cognitive science, at least as a starting point for scientific theorising.

When assessing whether the concept of belief should be eliminated from cognitive science, it is this generalised concept of belief that we should be considering, not folk usage. In this chapter I will consider the obstacles that we face in delineating BELIEF from folk usage of 'belief' and end by setting out BELIEF's functional role.

Smith (1982, p.505) notes:

> "The surface structure of our complex common-sense psychological theory is extremely complex. If we are to command a perspicuous overview of the theory, then it seems that we must try to isolate an underlying skeletal structure of psychological explanation which we can then overlay with a variety of finer discriminations, explanatory epicycles and so forth."

But isolating that skeletal structure is not so straightforward. In abstracting BELIEF from everyday usage there are at least three major obstacles. Firstly, the term 'belief', when used in certain contexts, has acquired several specialised usages in the English language. Secondly, the concept is often implicit in our sentence structures, and regularly remains unverbalised in everyday conversations. Thirdly, 'belief' is not the only term that English speakers use to express BELIEF in explaining and forming expectations about human behaviour. We use a variety of what I will call 'belief-like' terms – we can *believe that*, but we can also *hope that*, *suppose that*, *think that,* etc. These terms, I argue, are formed from the concept of belief in combination with other concepts to perform specific communicative functions.

From this complexity, philosophers such as Davidson (2006 [1963]) have identified the core of what they call **belief-desire psychology**, the idea that all intentional action can be rationalised by a belief-desire pair. The concept of belief that features as one of the pair in belief-desire psychology is not the 'belief' of folk usage, but rather the doubly theoretical concept of belief.

In abstracting and generalising the concept of belief from the complexity of everyday usage, philosophers have delineated a concept that is more fundamental to human mental understanding abilities than the details of how the word is used in everyday speech. But by repurposing the existing term 'belief' to use as a label for BELIEF, philosophers have created a degree of tension between the everyday and the philosophical usage of the term.

The concept that is to be integrated into science then, is not a confused and multi-use pre-scientific concept that is to be read from the folk usage of the word 'belief', rather it is already a doubly theoretical term, something far more suitable for an initial integration into cognitive science and a good start for scientific theorising. This, of course, it not to say that the concept will remain unchanged. Divergence of the scientific concept from its initial starting point is to be expected as scientists revise it to meet their purposes. I must be very clear that in this chapter I am not arguing, with Fodor, that belief and other folk psychological concepts will ultimately be vindicated by cognitive science, merely that it is a useful starting point.

In section **2.1** I will examine the possible constraint on the methodology of looking at English language usage only. In **2.2** I explore 'belief' and BELIEF in folk usage, focusing on the three obstacles to abstraction that I mentioned above. In **2.21** I look at how 'belief' functions in everyday language, in **2.22** I show how BELIEF is often left unverbalised, and in **2.23** I look at belief-like terms in which BELIEF is combined with some other concept to perform a specific communicative function. In section **2.3** I move on to philosophers' BELIEF and how it functions in belief-desire psychology and end by describing the functional role of the generalised concept of belief.

2.1 constraints on methodology – is BELIEF pancultural?

I will be arguing that the philosopher's concept of belief is doubly theoretical. Part of this is that the concept of belief is an abstraction from the way that people use the term 'belief' and other belief like-terms. Due to the constraints of space and my linguistic ability, I will be exploring usage of the term 'belief' in English only. It could be argued that what I am delineating is the concept of belief in English-speaking cultures. But I do believe that there is the potential for the results of this inquiry to be generalised.

There are clearly substantial variations between cultures in the 'surface structure' of people's reasoning about other people's thoughts and behaviour. Obviously different languages will use different words, but differences can go deeper than that. Take the example from the work of Penelope Vinden (1996), who studied children of the Junin Quechua people, who live in the Peruvian Andes. Junin Quechua do not have words that refer directly to mental states, they use for example, "what would he say?" rather than "what would he think?" (Vinden, 1996, p.1708).

Mental state words are also absent from many Junin Quechua folk tales. For example, the tale of a fox that, while standing on the shoreside, sees a large wheel of cheese floating in the middle of a lake. However, the narrator tells the audience that the cheese that the fox sees in the lake is not a cheese at all, but in fact the reflection of the moon. The fox, hungry for cheese and frustrated by the lack of assistance from a passing owl, sets out to swim to the centre of the lake to retrieve the cheese, but tragically drowns. While the story makes use of the appearance-reality distinction, it makes no reference to the mental states of the fox, something that western students include when they retell the story.

Despite their lack of mentalistic language, the hearers of the story clearly have an understanding of the fox's motivations. How else could they understand the motivation of the fox in taking the risk of swimming to the centre of the lake? They explain the fox's actions by saying that "it would say there is a cheese in the lake". But clearly they also assume that there is some kind of persisting state that is the cause of the behaviour, it is just that they refer to it only indirectly by its potential effects on verbal behaviour. It is likely then, that while the Junin Quechua lack words that directly refer to the mental state of belief, they do not lack the concept of belief. This is supported by the fact that in their everyday speech they often make use of Spanish loan words that refer to mental states.

Another theorist that emphasises cultural differences between folk psychologies is Lillard (1998). Reviewing ethnographic studies Lillard found that the folk psychologies of other cultures have several features that do not exist in what she calls "European American folk psychology". These features fall into four different categories:

1) Magical thinking, including the attribution of human psychology to spirits and extrasensory perception to individuals
2) Differences in the distinctions drawn between different mental states, mainly emotional states
3) Denial of negative states. For example, sadness in Tahiti is believed to leave one open to attack by evil spirits and so individuals often deny that they are sad, even to themselves
4) Different levels of value places on individuals' internal lives. Western thought tends to place a lot of emphasis on autonomous individuals and so consideration of internal lives is highly valued. In other cultures, consideration of others' internal lives is not so important. In Samoa, for example, blame is apportioned on the impact of one's actions, rather than the intention that caused it.

These differences all fall within what Lillard calls 'optional construals', that is they are variations where there is "no solid evidence one way or the other indicating the correct construal" (Lillard, 1998, p.5). In the terms that I am using, they are differences in surface structure, rather than at the core of folk psychology, and it is the core concept of belief that I am interested in.

Lillard doesn't rule out universal similarities in folk psychologies and suggests that one place that we could find them is by looking at how folk psychology develops in children. Henry Wellman and colleagues have undertaken extensive work looking at the differences in development of early mental understanding competencies across different cultures (see Wellman, 2014 for a review).

The developmental environment of Chinese and US children differs in several ways. Linguistically, the English word for 'belief' is neutral between the truth or falsity of that belief, whereas Cantonese and Mandarin have different verbs for thinking falsely. Culturally, in the US the focus is on autonomous individuals and a plurality of beliefs, while in China, there is more emphasis on shared knowledge (Wellman, 2014, p.98). Despite the differences in upbringing, children from both cultures showed a similar developmental trajectory in understanding that others can have beliefs that differ from reality.

In the past decade or so there has also been much work on non-traditional false belief tasks following from an original study by Onishi and Baillargeon (2005) who found that preverbal infants as young as 15 months showed some understanding of false beliefs. This suggests that some understanding of belief develops before full enculturation can take place.

Wellman (2014, p.40) argues that "These data clarify that culturally transmitted and culturally received information about persons and minds, indeed cultural learning itself, honours, in early development, some foundational understanding of persons and minds framed by a deeply mentalistic construal of persons, lives, and actions."

Key to any such foundational understanding will be a general concept of belief. I believe that if we dig far enough past cultural differences in surface structure and optional construals then we will arrive at a general concept of belief that is invariant across cultures. For the present chapter, however, I will limit myself to an exploration of usage in the English language, demonstrating that the concept of belief is invariant across cultures will have to wait until future work.

2.2 'Belief', BELIEF and the folk

In this section I will explore how the term 'belief' is used in everyday English language and identify three obstacles that we face when abstracting a generalised concept of belief from ordinary English language usage.

Before I embark on this task there is an important caveat to note. In much of this chapter I will be exploring everyday usage of the term 'belief' and other belief-like terms. We must remember that human communication is incredibly complex, certainly more complex than I represent it below. A difference in the tone of a speaker's voice or the subtle raise of an eyebrow can shift the meaning of a sentence substantially. However, this discussion is not intended as an argument that 'belief' and other belief-like terms have formal linguistic functions, or that certain belief-like terms always and exclusively imply the same thing. It is more an exploration of the variety of connotations that these terms can have in ordinary English usage and how this variation is achieved. I cannot hope that my discussion will be a comprehensive treatment. I do hope it will convince the reader that from the wide variety of usages and meanings that these terms have, philosophers have taken the common elements to delineate the core concept of belief, and that that core concept underlies the range of usages that I will explore.

2.21 'Belief' as it functions in language

Those not philosophically inclined would accept that, in general, the term 'belief' refers to a state that helps to guide the way that we behave. However, the way that the term functions in everyday usage is quite specialised.

There are at least three broad types of usage for the term. Firstly, 'belief' can function as a way of qualifying the endorsement of a proposition. Secondly, 'belief' can refer to a sense of confidence in an individual or collection of individuals. Thirdly, 'belief' can refer to beliefs about which there is disagreement and no hope for a straightforward resolution – for example, metaphysical beliefs, moral convictions or fundamental beliefs about the world or society.

I will tackle each of these senses in turn.

Something less than knowledge: 'belief' as endorsement qualifier

If someone is certain that a proposition is true, then the normal way of expressing this is by simply asserting that sentence. If you aren't certain that a proposition is true, then you should hesitate before asserting it if you don't want to mislead. One way of expressing your uncertainty is to add 'I believe that' to the start of the proposition.

Consider my friend Lee, who recently booked a trip to Poland[12]. He asks his cousin Simon what currency they use in Poland and Simon replies 'Poland use the Euro'. The following day, Lee visits a bureau de change and converts several hundred pounds Sterling into Euros for his forthcoming trip. Unfortunately for Lee, Poland use the Zloty not the Euro, a fact that he didn't discover until he arrived in the country. He had to convert his Euros to Zloty and had to pay additional commission for doing so. In this example, Lee would be quite justified in being a little angry with Simon. Simon answered Lee's question with the simple assertion that Poland use the Euro, which suggests that he was fully confident of the fact. However, Simon was not certain. To express this uncertainty, Simon could have responded to Lee's question with 'I **believe that** Poland use the Euro'. The addition of 'believe that' at the start of the assertion is acting as an endorsement qualifier – Simon would be effectively saying 'I think that X is the case, but I'm not sure.'

This endorsement qualifier would have changed Lee's behaviour. If, rather than simply asserting that Poland used the Euro, Simon had said 'I believe that Poland use the Euro', Lee could have taken this as a cue that Simon was unsure of his answer and he would have taken the precaution of checking the Polish currency before changing his money. In this case Lee wouldn't be justified in being angry at Simon. To Lee's 'it's your fault, you told me that they use the Euro!', Simon could rightly reply 'I only said that I **believe** that Poland use the Euro'.

The reason that belief functions in this way is because "I believe that P" is a weaker claim than an assertion of P. This creates the conversational implicature that the statement may be false. I shall unpack this below.

By asserting a proposition we are expressing a belief in that proposition. This is made clear by examples of what Wittgenstein called "Moore's paradox" (Wittgenstein, 2009, part IIx)[13]. For example:

"I purchased this jumper from the shop last Thursday, but I don't believe I did."

The two parts of the sentence refer to different things. The first part to a state of affairs and the second part to a cognitive state. But there is clearly something wrong about the sentence – one cannot coherently assert something and at the same time not believe it.

From the fact that we assert something as true, it follows that we believe it, so adding 'I believe that' to the start of an assertion does not on the face of it make the statement any more informative. In fact, this statement makes a weaker claim – "I believe that P" can be true even where P is false. This

[12] Example based on real events

[13] Although see Myers-Shultz & Schwitzgebel (2013) for what they see as potential counter examples to this.

additional and seemingly unnecessary information generates what Paul Grice (1975) called a conversational implicature (the thing that is implied is the implicatum).

Implicature in general is meaning or suggestion over and above what is strictly said. Conventional implicature is generated through the meaning of words, but conversational implicature is generated by exploiting the structure of discourse.

Grice noted that conversations, even very casual ones, are to some extent cooperative efforts between interlocutors. Because of this, we all have expectations that the people we are talking to will follow the Cooperative Principle – "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." (Grice, 1989, p.26). Beneath the Cooperative Principle, Grice identifies the four categories of Quantity, Quality, Relation, and Manner, and in turn each of the categories contain several conversational maxims.

In general, we expect that conversational partners will adhere to the Cooperative Principle and the four categories of maxims. Speakers can exploit these shared expectations to generate conversational implicature. Where one of the maxims is flouted, that is, the speaker clearly does not adhere to one of the maxims and is obviously making no attempt to mislead, then the hearer will attempt to square what has been said with the assumption that the speaker is still adhering to the cooperative principle and the conversational maxims. One of the ways of making what is actually said consistent with the speaker still adhering to the maxims is to attribute an implicature to the utterance.

Relevant to our present purposes are the conversational maxims that fall within the Quality category, namely (from Grice, 1989, p.26):

1) Make your contribution as informative as required (for the current purposes of the exchange).
2) Do not make your contribution more informative than is required[14].

By adding 'I believe that' to the start of an assertion, a speaker is flouting the second maxim of quality. Not only is this form of words explicitly stating that you believe the proposition – a seemingly unnecessary detail – it makes the statement weaker.

The hearer, in order to try to square this flouting with the speaker continuing to adhere to the cooperative principle and the maxims, interprets this as implicature. If the speaker was certain that the proposition P was true, they would have simply asserted P. By making a weaker claim, 'I believe that P', which is consistent with P being either true or false, the implicatum must be that they are uncertain of the truth of P.

The flouting of this conversational maxim can be made obvious by adding 'I believe' to the start of an assertion of any fact of which we would expect someone to be certain, such as assertions of shared common knowledge, autobiographical assertions, or assertions of perceptually obvious facts. Consider the following first-person examples:

"I believe that Boris Johnson is the current Prime Minister."[15]

[14] Grice was unsure whether the second of these two maxims was indeed a separate maxim or whether the same effect was already covered by the maxim within the category of quality: "be relevant" (p27). Arguably, adding 'I believe that' to the start of an assertion could also flout one of the maxims within the category of Manner – 'avoid obscurity of expression'. While this is potentially an interesting topic for future work, which and how many of the maxims is flouted by the addition is not important for our present purposes.

[15] True at time of writing.

“I believe that my first name is Andrew.”

“I believe that that tree over there is taller than me.”

Anyone I spoke to would assume that any of the facts contained in the above statements would be obvious to me. I follow the news, if anyone knows what my first name is it should be me, and that tree over there is the tallest in the park and quite clearly would tower over me. If I did utter any of these statements, then my interlocutor would be initially confused and quite probably explicitly wonder why I had phrased things in such a way as to make a weaker claim than would seem to be warranted. The way in which they would try to make sense of what I said as being consistent with the Cooperative Principle is to interpret me as trying to generate an implicature that all is not what it seems. Perhaps I am trying to suggest that while I believe that Boris Johnson is the current Prime Minister, there is a real possibility that certain conspiracy theories are true, and that the real powers of government are in the hands of others, lizard-men for example.

The same effect is achieved in the second and third person and again is most obvious where a sentence expresses a fact of which we would expect someone to be certain.

“You believe that your name is George”

“Sarah believes that she lives in England”

The implicature here being that George’s name is not actually George (although he believes it is) or that Sarah doesn’t actually live in England (although she believes she does).

In these examples I have used assertions that couldn’t reasonably be doubted to show how unusual it would seem to add ‘I believe’ at the start. In practice, the implicature needed to square the above statements with me continuing to follow the cooperative principle are so fantastical, that the most sensible thing would be to ask for clarification. But in more everyday cases such as the example of Lee, Simon and the currency of Poland, if Simon had said “I believe that Poland use the Euro” the implicatum is clear, while he believes that the Euro is the currency of Poland but is not sure.

Of course, in most ordinary conversations when a speaker adds ‘I believe that’ to the start of an assertion they are not consciously calculating that it will have such an implicature. Nor is the hearer actively inferring it from the speaker’s words. Adding ‘I believe that’ has long become conventionalised as a way of qualifying endorsement. We can and do use novel conversational implicature all the time, however. A less conventionalised example that we might expect to hear in an everyday conversation would be:

Sylvie: “Did you cook Hans that risotto last night?”

Celeste: “I *tried* to cook Hans that risotto last night”

We can accomplish many things without trying, you can impress someone without trying, for example. But, unless we are talking figuratively, cooking a risotto is not something that one does without trying. By stating that she *tried* to cook the risotto, a statement consistent with both successful and unsuccessful completion of the task, Celeste weakens the claim. Why would she do this? To generate the implicature that she had tried to cook the risotto, but had in fact failed in some way. This could be either a complete failure with the meal ending in the bin, or a partial failure with the meal tasting less flavoursome than expected.

In everyday usage then, the term ‘belief’ can be used by a speaker to qualify their endorsement of a proposition, to communicate to the hearer that while they may currently believe the proposition in question, they recognise the fact that they may indeed be mistaken. This is one usage of the term, but there are at least two others.

Systems of 'belief'

On learning that I am a student of philosophy and that my research is focused on 'belief', many people tend to assume that this involves the study of systems of belief, such as religion, non-religious moral beliefs or conspiracy theories. This is perhaps because people often assume that 'belief' in this sense is the most obvious subject for philosophical inquiry, but it also highlights a further way in which the term 'belief' is used in everyday language. 'Belief' is often used to refer to that specific class of beliefs that are not directly warranted by perception, that are optional (in that one would be able to function reasonably normally day to day without them), and around which groups of adherents or subcultures are likely to develop.

We can make this clear with an example. Hans is introduced to Celeste, they get on well. Hans offers Celeste some oysters, but Celeste declines.

Celeste: "I can't eat those, sorry."

Hans: "Why not?"

Celeste "Because of my beliefs."

What kind of 'beliefs' is Celeste likely to be referring to here? If Hans were to ask for more detail, he would perhaps expect her to talk about her religious beliefs. Perhaps Celeste is forbidden to eat shellfish by whatever creed she subscribes to. Another likely response would be that she cannot eat oysters due to her beliefs about the moral worth of non-human animals. However, responses that Hans would not be expecting would include things such as:

"I believe that the kitchen in this restaurant is quite unhygienic."

"I believe that I am allergic to oysters."

Using the term 'beliefs' in the way that Celeste did clearly implies that she is referring to a moral or religious belief. The key differences between these types of 'beliefs' and the 'belief' about poor kitchen hygiene is that the truth or falsity of moral beliefs cannot be settled by something so straightforward as a government health inspection. They are to some extent an article of faith. Perhaps because of this, moral and religious beliefs are deeply held and are often key parts of a person's identity. It would seem relatively normal for someone to identify as a Christian or as a vegan. It would seem less normal for someone to identify as someone who believed that there was poor hygiene at their local seafood restaurant.

I 'believe' in you

The third usage of 'belief' that I will discuss is belief in an individual, collection of individuals or a cause. When used in this way, belief is less an attitude towards the truth of a proposition and more a sense of confidence in ability, or that something is the right thing to do. Some examples will help to make this clear.

"I believe in you" is the title of different songs by Kylie Minogue, Dolly Parton, Bob Dylan, and Michael Bublé. They are not using this word to denote belief in someone or something's existence, as it would mean if I said that "I believe in ghosts", but rather to express their confidence in the ability or moral strength of someone. If someone was having a crisis of confidence before a job interview, to try to encourage them you might say "I believe in you" as a shorter and more poetic way of saying "I believe that your skills and experience make you suitable for this role and that you will be able to demonstrate this during the recruitment process."

'Belief' can be used in a similar way with a collection of individuals, such as a sports team. In the build up to a key match, former manager of Derby County Football Club, Frank Lampard said, "I **believe** in the players"[16] Lampard followed this up with, "Belief is key and we have to take belief with us to Elland Road; going there believing that we can turn the result around". Here again 'belief' is being used as an expression of confidence. In this case, Lampard uses it to express that despite the fact his team were beaten by rivals Leeds in the first leg of a two-leg match, that he 'believes' that the team has the ability to win the second leg and turn the result around[17].

In a different way one can believe in a cause or course of action, meaning that one believes that it is the right thing to do. In a revealing interview, former UK Prime Minister David Cameron gave his thoughts on his successor Boris Johnson. When talking about Johnson and his backing of Brexit in the referendum, Cameron said: "The conclusion I am left with is that he risked an outcome he didn't believe in because it would help his political career." [18]

What Cameron means here is that Johnson didn't believe that, on balance, leaving the European Union was the right thing to do. He believed that it would not be beneficial for the country. Rather, Johnson backed the cause because he thought that it was a way for him to attain the power that he had always craved.

'Belief' then, has several different but related usages in everyday English discourse. Firstly, as a way of qualifying the endorsement of a proposition. Secondly as way of referring specifically to systems of belief, metaphysical or moral *beliefs,* and other beliefs that are not directly warranted by the evidence and are therefore liable not to be agreed upon by all. Thirdly, as a way of expressing confidence in someone or something, or belief in the value of a cause or the correctness of a course of action.

While this brief survey of usage is not completely exhaustive, it covers the main categories. These specialist linguistic functions are well established in everyday usage, but they represent unwanted noise when it comes to isolating our generalised concept of BELIEF. To delineate a concept suitable to be integrated into cognitive science, we need to ignore the unwanted connotations of the term 'belief' that we outlined above.

2.22 BELIEF is often left unverbalised

In the previous subsection we identified some connotations that the term 'belief' has in everyday language that we ought to ignore when delineating our generalised concept of belief. In this section we explore a different problem. Rather than unwanted connotations, here we look at the way in which because of how deep-rooted belief-desire psychology is in our everyday social cognition, many elements are left unstated. This means that the concept of belief can feature as part of an explanation or prediction of behaviour, without 'belief' or any other belief-like term being mentioned. This is something recognised by Davidson who says that "A primary reason consists of a belief and an attitude, but it is generally otiose to mention both." Davidson (2006 [1963], p.26). This is because once the belief or the desire has been spelled out, it is often a straightforward process for the hearer to infer the other of the pair. Verbalising the pair according to Davidson, would be unnecessary. This is true, but further to that there is the danger, especially in the third person, that spelling out the pair in full could potentially also create an unwanted implicature, changing the meaning of the sentence. Let's consider some examples.

Firstly, an example in the first person.

Clare bumps into Sarah at the local shops. Clare is carrying a small bottle of oil. The belief-desire pair that rationalises Clare's purchase of the oil is:

[16] https://www.dcfc.co.uk/news/2019/05/i-believe-in-the-players
[17] Lampard's belief was well founded, Derby won the match and took the tie 4-3 overall.
[18] https://www.politico.eu/article/boris-johnson-didnt-believe-in-brexit-david-cameron/

Desire – to stop her bicycle chain from squeaking

Belief – oil will prevent a bicycle chain from squeaking

If Sarah asked Clare why she had purchased the bottle of oil, a satisfactory answer would be: “to stop my bicycle chain from squeaking”. Clare responds by stating her desire, while the belief remains unstated. No problem, it is a simple inference for Sarah to complete the pair to rationalise the action. If Clare has purchased the oil because she wants to stop her bicycle chain from squeaking, then she must believe that the oil will prevent the chain from squeaking.

The inference works the other way too. If Clare had responded to Sarah’s query by stating a belief: “A bit of oil will stop my bicycle chain from squeaking” Then Sarah’s inference from the belief to the desire is again straightforward, Clare desires to stop her bicycle chain from squeaking.

The structure is the same in a third person example.

Hans calls round to Celeste’s house for a visit but Sylvie, Celeste’s sister, answers the door. Celeste is not in. Sylvie tells Hans that Celeste won’t be back for a while, she has set off cycling to the shop across town. The belief-desire pair that rationalises Celeste’s action are:

Desire – to purchase Kenyan coffee beans

Belief – The large shop across town sells Kenyan coffee beans (but the shop this side of town doesn’t)

Hans, as curious as ever, asks Sylvie why her sister has gone all the way to the shop across town, when she could have gone to the shop this side of town, a much quicker journey. A satisfactory answer to Hans’ question would be to state just the desire: “She wanted some Kenyan coffee beans”. Hans would be able to infer Celeste’s belief that she is able to buy the coffee beans from the shop across town, but not the shop this side of town. Similarly, Sylvie could have responded “Because she believes that that shop sells Kenyan coffee beans.” And Hans would have been able to infer Celeste’s desire for Kenyan coffee beans.

In both the first person and the third person case, as Davidson noted, it would seem unnecessary to spell out the belief-desire pair when the inference from one to the other is so straightforward.

When using similar examples to explain to one of my friends the concept of belief as it is understood by philosophers, he looked puzzled and said: “yeah, but that’s just obvious though isn’t it?”. Yes, it is obvious. Indeed, these inferences are so obvious that we don’t need to make them consciously. If Sylvie told Hans that Celeste had gone to the shop across town because she wanted some Kenyan coffee beans, Hans would not have to muse to himself that Celeste must believe that they sold the desired beans at the shop across town, that’s “just obvious”. It goes without saying. Hans would just accept a simple statement of Celeste’s desire as a satisfactory rationalisation of her action.

But what if Sylvie had responded by spelling out the full belief-desire pair? Why did Celeste go to the shop across town? “because she wants some Kenyan coffee beans and she believes that the shop across town sells them.” This statement clearly generates an implicature, it implies that Celeste believes that the shop across town sells the beans, but she isn’t sure that they do. The same implicature would not be generated if Sylvie had left out “believes that” and had rather responded “because she wants some Kenyan coffee beans and the shop across town sells them”. But the fact that the shop across town sells Kenyan coffee beans is not what partially rationalises Celeste’s action. This fact could be unknown to her – what is important in rationalising Celeste’s action is that she **believes** it.

In everyday language then, it is common for the belief element of the belief-desire pair rationalising human actions to remain implicit in sentences. In most cases it is “just obvious”, as Fodor (1987, p.3) says

"Commonsense psychology works so well it disappears." But the fact that it remains implicit obscures the underlying structure of our thinking about the causes of human behaviour.

Uncovering the underlying structure of our mental understanding abilities is not just a matter of ignoring complexities, it is also a matter of reconstructing what is so obvious to us that it remains unsaid.

2.23 BELIEF is often used in combination with other concepts

As we saw from our earlier discussion in subsection **2.21**, the term 'belief' has a specific specialist usage in everyday talk. This usage does not match closely with the philosophical concept of belief.

The closest that we get to a statement of belief in everyday language does not involve the term 'belief', it is the simple assertion. By default, an assertion will be taken as a statement of fact when it comes from a trustworthy source. For example - 'Brussels is the capital of Belgium', asserted by a speaker who is not a known liar and has a basic grasp of European geography, will be accepted as true by those that were previously ignorant of that fact. Even though the statement may be false, the assertion is our most straightforward way of communicating what we believe to be the case.

However, in many instances of everyday usage, a speaker may wish to communicate more than a simple belief. Tagging further information about the specific cognitive role that the belief has can help to guide the appropriate output of the hearer. Rather than spelling out several different pieces of information separately, the quickest and easiest way of doing this is by using belief-like terms.

As discussed above, the term 'belief' modifies an assertion, conveying more information than the assertion alone. In the case of 'belief' it suggests that the speaker is not fully confident of the truth of the statement being asserted. While 'belief' plays this role, there are many other belief-like terms that convey possession of BELIEF, but supplement it with further information.

Rather than using a single general term, common usage tends towards a multitude of more specialist terms and this acts to obscure the general concept from view.

In this subsection I will explore several belief-like terms and explicate how these different terms provide supplementary information to a basic statement of belief.

Specifically, I will explore belief-like terms that modify a statement of belief by:

- Qualifying endorsement
- Reporting on cognitive role
- Reporting on emotional attitude.

Endorsement qualifier

We saw earlier how the term 'belief' can be used to generate an implicature that there is uncertainty about the proposition. But 'belief' is not the only belief-like term that can play this kind of role and different belief-like terms can be used to express different levels of endorsement. Certain terms can indicate low confidence, others high confidence but with reservations, still others unequivocal endorsement. This additional information can then act as a guide for the hearer, influencing their subsequent behaviour. While a straight assertion might lead the hearer to simply act on the basis that the proposition is true, an assertion with an endorsement modifier could lead the hearer to take a different course of action, or to question the asserter further about their exact level of certainty, how they came to believe the proposition, and the reasons that they believe it to be true.

Terms such as '**sure that…**' and '**certain that…**' have the function of communicating high levels of endorsement. Thinking back to our example earlier of Lee, Simon and the Polish currency, if Simon had said that 'I'm sure that Poland use the Euro' then Lee's anger would again be justified. As it would be if Simon had used 'certain that…'. Both 'sure that…' and 'certain that…' indicate a high level of endorsement, but they are both sensitive to context. While in many situations they can communicate the full unequivocal endorsement suggested by the straight assertion, in some situations that doesn't seem to be the case. Compare the following two responses to the question: 'I'm hungry, do we have any food?'

'There are sausages in the fridge'

And

'I am sure that there are sausages are in the fridge'

The first sentence has full endorsement because there is no question whether the assertion is true. The second sentence, however, despite the assurance that the asserter is sure that the sausages are in the fridge, seems less certain to be true because it generates an implicature and raises the possibility of error. 'Sure that…' and 'certain that…' can have degrees. If you are sure that there are sausages in the fridge, how sure are you? If you are certain, how certain? In the above situation 'sure that…' is acting as a qualification, albeit a weak one, of the assertion. Confidence in the proposition is still high, but the door to possible error is left open – I'm sure there are sausages in the fridge, but I could be mistaken.

In contrast, '**know that…**' seems to function as a binary – you either know or you don't. 'Know that…' is often used to express unequivocal endorsement where the truth of the proposition has been questioned. Consider:

Alice – 'Are you sure these mushrooms are safe to eat?'

Sarah – 'I know that they are'

In this situation, Alice had better hope that Sarah knows, rather than thinks, believes, or is even just sure that those mushrooms are safe to eat.

The currency situation and the mushroom situation are quite serious examples. There is a high cost attached to error – time and financial loss for Lee and his Euros, poisoning and potential death for Alice and her mushrooms. In high cost situations such as these the endorsement modifier is highly likely to result in behavioural changes on the part of the hearer. But there are other examples where the stakes are lower and the hearer's behaviour may be unaffected by the modifier. 'I think it's going to rain' for example, said to Clare who is about to set out on a short journey may not result in her borrowing an umbrella just in case. In this instance, she judges that the effort required to return the loaned umbrella outweighs the risk of getting wet. If she did get wet however, she couldn't say that she wasn't warned.

As we have seen, an unmodified assertion of a proposition defaults to full endorsement. Belief terms can be used to qualify endorsement or to emphasise it where the truth of the proposition is questioned. Belief terms, for example 'think that…' indicate weak endorsement, while others such as 'sure that…' indicate strong levels of endorsement. Finally, some belief terms, such as 'know that…' can be used to express full and unequivocal endorsement of a proposition where the truth of that proposition is in question. The hearer may or may not alter their behaviour in light of the endorsement modifier depending on their judgements of the cost attached to error and the cost of taking precautions and/or removing uncertainty.

Report on cognitive role

In addition to signalling the level of endorsement of a proposition, belief-like terms can be used to report on the cognitive role of a proposition. This could involve revealing how the speaker came to endorse the proposition or showing how endorsement of that proposition is influencing the behaviour of the speaker. While the endorsement modifier discussed above comes with an embedded judgement, the report on cognitive role allows the hearer to make their own judgement.

Consider '**feel that…**', which is often used to indicate that a speaker's endorsement of a proposition is based on intuition rather than the result of a conscious inference. In these instances, the speaker may find it difficult to justify their belief to themselves and others[19].

For example, 'Feel that…' is commonly used in aesthetic, moral or value judgements – 'I feel that Trellick Tower is a more aesthetically pleasing building than Buckingham Palace' seems natural, as does 'I feel that killing one person to save five is wrong'[20]. I am not arguing here that aesthetic, moral, or value judgements are never based on conscious reasoning. You could, for example, argue that Trellick Tower is more aesthetically pleasing because its rugged functionality makes more of an emotional impact on the viewer than the safer and more traditional Buckingham palace. You could argue that killing one person to save five is wrong because it is never right to take a human life. But you could also just feel that Trellick Tower is more pleasing, or that killing one to save five is just wrong.

With the cognitive role of the proposition made clear, the hearer can judge how far they want to endorse the assertion. In the case of 'feel that…' such a judgement may be based on how far you share, for example, the aesthetic tastes of the asserter.

While 'feel that…' seems natural in cases of aesthetic, moral or value judgements, it doesn't seem natural in more factual cases, for example 'I feel there are sausages in the fridge' or 'I feel that Brussels is the capital of Belgium'. This is because these are not the type of things that one just 'feels that'. In these types of cases there are generally reasons that one can point to in order to support their endorsement. Perhaps Clare told you that she would be shopping for sausages earlier that day and the fridge is the place that she would naturally store them. Perhaps you watched a travel show where the host travelled around the Low Countries and you clearly remember that there was a parliament building in Brussels.

In contrast to 'feel that…', '**reckon that…**' implies that the assertion is the result of a chain of reasoning, albeit an imprecise one. The proposition could be intended as a rough estimate, for example, on a recent cycling holiday Chris said to me 'I reckon it's another five miles before we reach Rotterdam'. This assertion could be based on the fact that we cycled through a village that marked the halfway point an hour ago, or perhaps the environment is starting to become less rural and more suburban. As Chris has reported on the cognitive role of the proposition, I can make a judgement on how accurate it is likely to be. In this instance, I have no better information on which to base an estimate of distance and I know that Chris is generally a pretty good judge of these things so I accept it. As Chris has told me that he reckons we are five miles from Rotterdam I know that the estimate is imprecise, so if it turns out to be four miles or six miles I'd still be happy that it's a decent estimate.

In this way we can see that 'reckon that…' can play a dual role, as both a report on cognitive role and an endorsement modifier implying that the proposition should not be taken as precise.

[19] A complication here is the large body of literature which demonstrates that humans are poor at identifying the reasons behind their own decisions and judgements. Many subjects' explanations of their own behaviour can be revealed to be post-hoc rationalisations, see, for example, Wilson (2002). For the purposes of this paper, what is important is that 'feel that…' tends to be used where the speaker cannot offer reasoning behind their endorsement of a proposition, regardless of whether or not that reasoning was genuinely a factor behind the endorsement of the proposition.

[20] I am not arguing that 'feel that…' is the only or correct belief term to use in these cases – 'I think that Trellick Tower is more aesthetically pleasing building than Buckingham Palace' also seems perfectly natural. But these different belief terms would serve different communicative functions.

Another belief term that functions as a report on the cognitive role of a proposition is '**suspect that…**'. 'Suspect that…' is generally used where someone has limited evidence for an assertion, but still feels that it is a likely scenario. However, there is a social cost to being incorrect, so the asserter can't openly make the assertion. For example, Alice whispers to Sarah 'My biscuits have been disappearing very quickly. I suspect that Rob has been helping himself while I'm out of the office.' Alice has good reason to suspect Rob. He frequently works alone in the office and he is renowned for his sweet tooth. While this evidence is circumstantial, he has both motive and opportunity. By using 'suspect that…' Alice is communicating her lack of conclusive evidence to Sarah. 'Suspect that…' therefore acts as an endorsement modifier, but it also goes further than that, in hinting at a cognitive role – functioning to suggest the type of evidence that supports the assertion and more often than not, a reluctance to confront the suspect or make the assertion openly because of a potential social cost[21]. This is the case with Alice, she is reluctant to confront Rob because if she has got it wrong, he may be upset at the false accusation and this could sour the atmosphere in the office.

There is also an implication with 'suspect that…' that the thing suspected is to be disapproved of or is in some way negative. It fits well with the above example of biscuit theft, but it doesn't seem to quite fit as well with, for example, Alice suspecting that Rob is a generous charitable donor or suspecting that Rob is a talented pianist. In this way, 'suspect that…' operates as both a report on cognitive role and a report on emotional attitude, something that we will explore in the next section.

A further example of how belief like terms can act as reports on cognitive role is '**surprised that…**'. This term can imply that the asserter held a certain belief or set of beliefs, but has subsequently received evidence that has caused them to update that belief or set of beliefs. For example, while watching a nature documentary, Clare states 'I'm surprised that a whale is a type of mammal'. Clare's statement suggests that she currently believes that whales are mammals, but that she once thought that they weren't. She is communicating that her previous belief was contradicted and has now been updated.

Clare's statement of surprise need not imply that she had previously explicitly held a contradictory belief, say, that whales were a type of fish. In fact, Clare need not have ever thought about which class of animals the whale belongs to for her to be surprised by the fact that that they are mammals. It wouldn't be contradictory for her to say that "I've never really thought about what class a whale was in, but I'm surprised that they are mammals." It may be that, based on her previous set of beliefs, when prompted she would have guessed that whales belonged to a class of animals other than mammals.

As we have seen from the examples of 'feel that…', 'reckon that…', 'suspect that…', and 'surprised that…', everyday belief terms can go further than just suggesting differing levels of endorsement of an assertion. These examples show that belief-like terms can function to communicate something of the cognitive role that a proposition is playing for the asserter. In the cases of 'feel that…' and 'reckon that…', the belief-like term suggests that a certain type of reasoning or evidence is causing the belief. The case of 'suspect that…' additionally hints at the future behaviour of the asserter, i.e. that they will be reluctant to openly assert the proposition. The example of 'surprised that…' gives some indication that the speaker previously held different beliefs and that they have now been updated.

Other examples of belief terms that communicate cognitive role are **'imagine that…'**, **'suppose that…'** and **'anticipate that…'**.

Emotional Attitude

The final class of everyday belief-like terms that I will discuss are those that package a belief with an emotional attitude towards the truth of the proposition. In this way they communicate the feelings of the

[21] 'Suspect' obviously has a technical legal usage, which is clearly related to the more informal usage of the term that I focus on here.

asserter towards what is being asserted. These types of belief terms can guide the reactions of hearers and can also suggest the future behaviour of the asserter in light of the truth of the proposition.

For things that have happened and that we are certain of – terms such as **'happy that…'** or **'glad that…'** suggest that the state of affairs was desirable. For undesirable states of affairs, we would use terms such as '**unhappy that…**' or '**sad that…**'. Terms like this can go some way to guiding the expected reactions of hearers. For example, 'Congratulations!', would be an inappropriate and somewhat irksome response if someone informed you 'I'm unhappy that my company is relocating and my job is moving to Belgium'. In this example we could also predict the future behaviour of the asserter – as they are unhappy that their job is moving to Belgium we can assume that they will be reluctant to move to keep their job, and that they may start looking around for alternative jobs closer to home. For things that may or may not happen in the future, we can express our emotional attitude towards the state of affairs described by adding '**hope that…**' or '**fear that…**'.

As previously mentioned, the above is not intended as an exhaustive discussion of the various connotations of the belief-like terms that we use in everyday language. They can provide information about how far the asserter endorses their belief, what cognitive role the belief is playing, and the emotional attitude towards the state of affairs the belief is about. This additional information can communicate the likely future behaviour of the asserter and/or inform the future output of the hearer. The emotional attitude expressing terms discussed above even merge into a single term encompassing the functions of belief and desire, two clearly distinct states in the idealised model of belief-desire psychology. While belief-like terms clearly function in a far more complex way than the concept of belief functions in an abstracted belief-desire psychology, I believe the above discussion shows that the additional complexity is due to the concept of belief being packaged with a cognitive report, which may even include elements of desire. Since these components are dissociable, then abstracting belief-desire psychology and presenting it as the core of everyday folk psychology seems entirely justified.

As I have demonstrated in this section, in English there are three major hurdles for thinkers looking to delineate the core which underlies our mental understanding abilities. Firstly, we have specialist uses of the term 'belief' and the purest expression of a belief, the straight assertion, contains no reference to a mental state at all. Secondly, we have the fact that we more often than not don't fully verbalise the underlying structure of our mental understanding abilities. Finally, we have a variety of belief-like terms that contain the general concept of belief but either package the concept up with other concepts, or have specialist connotations or implicatures. It is from this complex surface structure that philosophers have abstracted the underlying structure of belief-desire psychology.

The central claim of this chapter then is to show that belief is a doubly theoretical term. BELIEF is a theoretical concept firstly in that it plays a functional role in a theory of how the human mind works, a theory that everyone uses daily to understand and form expectations about human behaviour. But BELIEF is theoretical in a second sense – philosophers have worked to delineate the concept from a complex surface structure of usage in which it is far from readily apparent. Rather than simply lifting the concept from folk usage of the term, philosophers have reconstructed the concept through a process of abstraction and generalisation. In the next section, I will explore how that abstraction proceeded, and how the generalised concept of belief is given meaning by its functional role in the core theory of belief-desire psychology.

2.3 Philosophers' BELIEF

In the previous section we looked at some of the difficulties of isolating the concept of belief from the complexity of English language usage. We looked at the specific usages of the term 'belief' in the English language, the fact that BELIEF is often left unverbalised in ordinary language explanations of behaviour, and how BELIEF is packaged up with other information in belief-like terms. In this section, I will focus on the idea that, at the core of this complexity, the most basic structure underlying our mental understanding abilities is the interaction between two types of state – beliefs and desires. The concept of belief then, is functionally defined in this core belief-desire theory of how these two types of state interact with input and each other to produce output. I will finish the chapter by setting out a broad account of the concept of belief that forms one of the two central states of this belief-desire psychology.

The origin of the idea that people use theoretical mental state concepts to explain and form expectations about others' behaviour is perhaps found in the essay 'Empiricism and the Philosophy of Mind' by Wilfrid Sellars (1997 [1956]). Sellars' aim was to set out the resources that would need to be added to a behaviouristic language to allow people to understand seemingly intelligent behaviour where the agent in question was not giving a running verbal commentary of their actions. To explain, Sellars developed a myth in which our "Rylean ancestors" (Sellars, 1997 [1956], p.102) could only use terms that described overt behaviour. Into this mythical world came a man named Jones. Jones invented a theory in which overt behaviour (including verbal behaviour) is "…the culmination of a process which begins with certain inner episodes" (Sellars, 1997 [1956], p.103). These discursive inner episodes Jones called 'thoughts'. Through the myth of Jones, Sellars introduced the idea that we understand human behaviour by modelling it using a theory that posits several different functionally defined unobservable states that interact with each other to produce observable human behaviour.

While Sellars introduced the idea that we used a theory of the human mind, he didn't go into details about the different theoretical concepts involved. One early and influential paper on the role of the theoretical state of belief in understanding human behaviour is Donald Davidson's 'Actions, Reasons, and Causes' (2006 [1963]).

What I have been broadly calling mental understanding abilities Davidson frames as rationalising human actions. When we give a reason for an agent's action, we rationalise that action. Davidson's primary aim in his essay is to argue that rationalisation is a type of causal explanation. In order to do that he takes care to outline his idea of a **primary reason** for an action.

The first thing to note is that reasons for human actions are clearly very diverse. Any single action could have a wide range of different rationalisations. Take something simple, like Hans jogging to a friend's house. Straight away we can think of at least two different potential reasons why he would choose to jog rather than walk or use some other mode of travel. Firstly, he may have wished to reach his destination more quickly. Perhaps he had arranged to visit at 8pm but had set off later than planned and needed to make the time back. Secondly, Hans may be a fitness enthusiast and he could have been using the opportunity of travelling to his friend's house to get in a little cardio work. We could think of many more reasons that could rationalise Hans' decision.

There are an indefinite number and variety of reasons that could be used to rationalise the full range human action, so if Davidson is to characterise a reason for action, he needs to do this at a high level of generality. He needs to identify, beneath all the diversity, the elements that all reasons for action have in common.

Davidson sets out two necessary conditions for something to be a reason for an agent's actions (from Davidson, 2006 [1963], p.25):

> C1. R is a primary reason why an agent performed the action A under the description of d only if R consists of a pro attitude of the agent towards actions with a certain property, and a belief of the agent that A, under the description of d, has that property.
>
> C2. A primary reason for an action is its cause

Focusing on C1, we can see that for Davidson, all primary reasons for action consist of a pairing of what he calls a ‘pro attitude’ and a belief.

On this account, what rationalises Hans’ action of jogging to his friend’s house? On my first rationalisation, the pro-attitude is being on time and the belief is that jogging will result in him getting to his destination more quickly. On my second rationalisation the pro attitude is maintaining his fitness and the belief is that jogging will achieve this.

Davidson’s primary reasons describe our folk belief-desire explanations of behaviour well, even though, as we saw in section 2.22, both the pro attitude (what I have been calling a ‘desire’) and the belief can often be left unverbalised.

I mentioned above that if Davidson is going to characterise all reasons for human actions he needs to do so at a high level of generality. One obvious generalisation in the first condition is the ‘pro attitude’. This is a technical term that covers a variety of different states which dispose the agent positively towards a state of affairs, examples of which include “…desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agent directed towards actions of a certain kind.” Davidson (2006 [1963], p.23).

Think of the diversity of pro attitudes that could rationalise someone joining the Navy. It may be that for Alice, she had a pro attitude for an enjoyable career in which she would travel extensively, meet new people, and learn new skills that would enable her to prosper in whatever career she chose to join after she left. This pro attitude, paired with the belief that the Navy could provide this type of career for her, formed her reason for joining.

For Sarah, her pro attitude was to please her parents. She believed that joining the Navy would achieve this because for three generations prior her family had been naval officers and it was expected by relatives that Sarah would follow the same career path. While the weight of expectation forms part of the reason for her joining, she may not be looking forward to starting her new career as she is a pacifist and is scared of water.

Despite the clear differences, Alice’s hope for an enjoyable career and Sarah’s desire to keep her parents happy would both fall under the category of ‘pro attitude’.

Pro attitudes can also vary in intensity. In commenting on different strengths of pro attitude, Davidson notes that, “Wanting seems pallid beside lusting, but it would be odd to deny that someone who lusted after a woman or a cup of coffee wanted her or it.” (Davidson, 2006[1963], p.25). To lust after something would be a particularly strong pro attitude, it may be that Rob has been lusting over a slice of cake all morning, so much so that he has had a difficult time concentrating on his work. He constantly checks the clock, counting down to his lunch break. Discussions with colleagues that morning revolved around cake and the relative merits of different types of cake. Simon however, had been laser-focused on his work that morning, but during a brief moment of pause thought to himself “I might treat myself to a slice of cake at lunch”.

Despite the clear differences in intensity of Rob and Simon’s desire for cake, both can be characterised as having a pro attitude towards cake, and that pro attitude could form part of the reason for each of them to make their way to the local patisserie during their lunchbreak.

As the previous example suggests, as well as varying in intensity pro attitudes can vary in duration. Clearly Rob has been lusting after cake for several hours, whereas Simon may have just thought about cake a few minutes before his lunchbreak. But the variation can be even greater than this. Sarah, may have known that she was expected to join the navy from a young age. Alice, however, throughout her life has been unsure of which career to pursue after she finished her education, and was only attracted to a naval career several months before she did so after talking to a naval recruiter at a careers fair.

Pro attitudes can arise through a whim or a sense of duty, they can be intense or they can be tepid, they can be fleeting or they can be held all of one's life. But Davidson is not interested in the diversity of the states, his task is to delineate the underlying structure of all reasons for action. To do this, he abstracts from this diversity what he sees as the common element shared by all the states he groups together under the heading of pro attitude – they are states that motivate an agent to bring about a particular state of affairs. The relationship between lust and pro attitude can be compared to the relationship between species and genus.

While Davidson goes into some depth about the diversity of states that fall under the heading of pro attitudes, he does not go into as much depth about the generalisations behind the other state in the pair of states that form a primary reason – BELIEF. As a label for this second category he uses 'belief', a word that plays a role in everyday talk about the behaviour and the inner lives of our fellow humans, but as we established in the previous section the term can also play a specialist role. Davidson however, intends the term to cover a concept that plays a generalised role that encompasses many belief-like states, what I have been referring to as the concept of belief or BELIEF.

Davidson chose the technical term 'pro attitude' rather than an existing term such as 'desire' because, as discussed above, 'pro attitude' is intended to have a much more general application than 'desire' has in ordinary language. The term 'desire' also has unwanted connotations. It suggests a strongly felt and enduring pro attitude. It doesn't seem suitable to describe, say, a sudden whim for a walk around the block to clear one's head.

Similarly with the term 'belief'. I have already discussed the specific ways in which 'belief' functions in everyday language, the connotations it has and the way in which it is used to generate conversational implicature. So why didn't Davidson choose a technical term for BELIEF, rather than sticking with the existing folk term 'belief'?[22]

One reason that Davidson may have been satisfied with the term 'belief' to stand for the core state underlying our rationalisations of human behaviour is that 'belief' has had a long history in epistemology of being used to refer to a fundamental generalised state that is prior to knowledge. It is a fundamental state, for example, in the traditional analysis of knowledge as justified true belief (from Ichikawa & Steup, 2018):

> *S* knows that *p* iff:
>
> 1) p is true;
>
> 2) S believes that p;
>
> 3) S is justified in believing that p.

According to this analysis I can, for example, be said to know that there is coffee in the cupboard if I believe it, I am justified in believing it (I put it in there myself 20 minutes ago), and there is indeed coffee in the cupboard. Justified belief falls short of knowledge however, if p is not true. If unbeknownst to me Clare took the coffee out of the cupboard 10 minutes ago then, even if from my

[22] Interestingly, Davidson's technical term 'pro attitude', hasn't quite caught on. When discussing basic cognitive architecture many authors use the term 'desire' in a technical generalised sense, rather than Davidson's term 'pro attitude'. For example Peter Carruthers' (2006 p66) basic cognitive architecture, which he argues is likely shared by all animals including some invertebrates, involves perceptual and bodily states feeding into 'desire generating' and 'belief generating' systems. Most of the time I use the terms pro attitude and desires interchangeably in this thesis.

point of view I have the justified belief that there is coffee in the cupboard, then I cannot know that there is coffee in the cupboard, because there isn't any coffee in the cupboard. Similarly, true belief falls short of knowledge. I can't be said to know there is coffee in the cupboard if on a recent visit to a psychic she told me that coffee would appear in my empty cupboard and I believe her. Even if, during my visit to the psychic, Clare took a trip to the shop to purchase some coffee and put it in the cupboard. I can't be said to know that there is coffee in the cupboard, because my belief is not justified. Even if, as is the case, there is indeed coffee in the cupboard.

The belief element in this analysis is not 'belief' as it functions in everyday language, but rather the generalised BELIEF that I have been discussing in this chapter. Whether I believe there is coffee in the cupboard is revealed by whether my output can be rationalised by my believing that there is coffee in the cupboard. If when asked whether there is coffee in the cupboard I reply in the affirmative, if when I am moved to make a cup of coffee I open the cupboard and reach inside, or if I infer that I have roast beans in the cupboard, then I can be said to believe that there is coffee in the cupboard. I have a representational state that causes this characteristic output. It is important to note that my belief causes this characteristic output whether the belief is justified or not and whether there is coffee in the cupboard or not.

One noteworthy feature of belief is the way in which two people can have the same belief with different justifications. Both Celeste and Hans share the belief that there will be heavy rain on the 16th day of the month. Celeste believes this because she is a student of meteorology and she has been tracking a weather front moving in from the Atlantic for several days. Hans, however, believes there will be heavy rainfall because it his birthday on the 16th of the month, and in his experience there is always heavy rain on his birthday (despite there being several historical exceptions to this). Both Hans and Celeste share the belief that there will be heavy rainfall on the 16th, and that shared belief will guide their behaviour in the same way – they both bring an umbrella for their planned walk in the park, but their reasons for possessing their belief are very different.

This philosophers' concept of belief, like the pro attitude, is an abstraction from the variety of belief-like states, identifying the core common to all of them. Like pro attitudes, beliefs can vary in strength and duration. The category of belief can contain both the lifelong and deeply held belief in the existence of a supreme being and the belief that there are coffee beans in the cupboard formed after me placing them there five minutes ago. As we saw in section 2.23, by using belief-like terms, we can package the core state with additional information such as how far we endorse the belief, how we came to have the belief, or our emotional attitude to the belief. While 'belief' is a common term in English, the straight assertion of X is a closer expression of our belief that X, than explicitly stating "I believe that X".

There is one more thing worth noting before I set out my general account of the concept of belief, because beliefs represent a state of affairs they can be correct or incorrect. Beliefs are aimed at correctness, as having true beliefs will generally result in more effective behaviour. At least this is certainly true in the case of immediately practical beliefs, for example about the location of objects. If I want to navigate to the shop, I am best served by a belief that accurately represents the location of the shop. But as an aside, arguably pragmatic considerations rather than truth can be the most important factor behind their adoption under some circumstances. I am thinking of people who have escaped from destructive cycles of behaviour by being 'saved' by religion or sportspersons being driven to great performances by a belief that they can win which contradicts analysts and bookmakers.

Returning to the general account of belief's functional role, a belief is a state that:

1) Has content that represents a current, past or future state of affairs

and

2) Has a functional role such that:

> 2.1) on the input side it is characteristically caused by perception, testimony or by inference from other beliefs. It is sensitive to information about its content, so that it is liable to be revised, weakened or cease altogether on presentation of contrary evidence, and is liable to be acquired or reinforced on presentation of positive evidence.
>
> 2.2) on the output side it combines with desires and other pro-attitudes to guide conduct, it combines with other beliefs to produce inferences and expectations, and also produces attitudinal change related to its content.

This functional role sets out how the concept of belief interacts with input, other beliefs and with pro attitudes to produce the variety of human output. This functional role, taken together with a complementary role that could be constructed for a pro attitude or a DESIRE concept, forms the basic conceptual framework of belief-desire psychology. According to this account of BELIEF, our mental understanding abilities are achieved by using this conceptual framework as a theoretical model of the cognitive architecture that underlies human behaviour. The functional role of belief describes how the state functions within this theory.

To be clear, I am not claiming that philosophers are in unanimous agreement that this is the correct account of the concept of belief. Indeed, Davidson (1974) would later argue that in order for an agent to have a belief, that agent must have the concept of belief. And that to have a belief one must be a language user. But for the purposes of integration, this type of restriction is undesirable. It is better to take the work that philosophers have done to identify the basic structure underlying our mental understanding abilities and let the cognitive scientists determine if the general concept needs to be refined, whether that be by adding further linguistic restrictions to its possession or fracturing it into sub-states.

We can see that this kind of generalised treatment of the concept of belief has been influential amongst contemporary philosophers and that belief features as a key component in their cognitive architectures. Peter Carruthers, for example, sets out a basic cognitive architecture which features beliefs and desires feeding into an action planning system. He argues that this basic architecture "…is likely to be shared by almost all animals that possess some sort of central nervous system." (Carruthers, 2006, p.66). While Nichols and Stich assume a "representational account of cognition" (Nichols & Stich, 2003, p.14) with a basic belief-desire architecture. Nichols and Stich also make no assumptions about the form of the representations underlying beliefs, an approach that I agree with and will discuss further in the next chapter.

While my generalised account of belief is not directly contradictory to the functional role described by Schwitzgebel that I discussed in chapter **1**, my account stresses the generality of belief, whereas the Schwitzgebel role seems more aimed at a specific type of verbal and consciously accessible state. Consider points 1) and 4) (from Schwitzgebel, 2011):

1) Reflection on propositions (e.g., [Q] and [if Q then P]) from which P straightforwardly follows, if one believes those propositions, typically causes the belief that P.

4) Believing that P, in conditions favoring sincere expression of that belief, will typically lead to an assertion of P.

Point 1) suggests that beliefs are about propositions and are available for reflection, while 4) suggests that they are verbal, or at least verbalisable. While, as I noted earlier, the use of the word “typically” means that the functional role does not rule out non-verbal, non-propositional beliefs, or beliefs that we can’t reflect upon, Schwitzgebel’s functional role does seemed to be aimed specifically at traditional verbal belief, rather than the generalised, doubly theoretical state that I am describing. Indeed, a more general description of the functional role of belief such as the one above would not have led to Jenson’s (2016) erroneous eliminative argument based on the fragility of the concept.

One thing that my account does have in common with Schwitzgebel’s is that it is not intended as a strict definition. My intention is not to set out a list of necessary and sufficient conditions for something to be a belief. Rather, the generalised account of belief is more descriptive, giving an indication of the range of functional roles that beliefs can have.

I have shown in this chapter that philosophers’ concept of belief, as it is set out in the generalised account of belief described above, is not merely a restatement of a folk concept. Rather, it is a doubly theoretical concept. It is a theoretical concept firstly in the sense that it is not simply read off from folk usage, which as we saw in section 2.2 is complex and jumbled, but instead is an abstract reconstruction of part of the basic conceptual framework that underlies human mental understanding abilities. It is also theoretical in the second sense that this basic conceptual framework is theoretically structured and that the concept of belief is functionally defined by its role in this theoretical framework.

The concept of belief then, as it features in the basic framework of belief-desire psychology, is a far more suitable candidate to be integrated into cognitive science than any old folk concept. In this thesis my aim is to determine whether the concept of belief should be eliminated from cognitive science. In this chapter I have shown that in answering that question, we must consider the generalised concept of belief rather than the varied and multi-faceted way that ‘belief’ functions in everyday discourse.

3. What is a *belief*?

In chapter 2 I developed an account of the concept of belief as a concept that plays a functional role in a theory of how the human mind works. Such an account construes belief realistically, the theory works because there is a type of state in the mind that realises the functional role described in the theory. For an agent to have a belief is for that agent to possess a representational state, a *belief*, in their mind that plays the functional role described by the generalised account of BELIEF. This type of account, in which belief possession is determined by internal factors, is known as **representationalism.**

Such an account leaves the concept of belief open to elimination. Once the theory is integrated into cognitive science, it is open to revision and refinement. As we saw in chapter 1, one potential refinement is the fracturing of BELIEF into two or more sub-states, leaving the fractured concept of belief redundant and without explanatory value. If this were the case, then BELIEF should be pragmatically eliminated from cognitive science. While redundant for scientific purposes, the concept could continue to play an active role in folk discourse.

Representationalism is not the only account of what it is to have a belief, however. Rather than the conditions for having a belief being possession of an inner representation, some philosophers suggest that to have a belief is rather to be disposed to exhibit a certain pattern of behaviour. On this type of account belief possession is determined by external rather than internal factors and is known as **dispositionalism.**

The dispositionalist account provides a potential way of avoiding Jenson's (2016) eliminativist challenge. As we will see in 3.32, the philosopher Eric Schwitzgebel argues that his dispositionalist account is able to handle what he calls cases of 'in-between belief' where it isn't clear whether an agent has a belief or not (Schwitzgebel, 2002; 2003; 2013). Where Jenson argues his empirical examples show that different detection methods yield different results, demonstrating that the concept of belief is fragile, the dispositionalist could claim that the agent rather 'in-between believes'.

In this chapter I will weigh up the relative merits of representationalist and dispositionalist accounts of what it is to have a belief. Once we have done so, we will be better placed to evaluate the eliminative challenge.

Below, in **3.1** I will set out how the representationalist account combines with my generalist account of the concept of belief. In **3.2** I will set out in some detail the traditional dispositionalist account and the more recent phenomenal dispositionalist account of belief put forward by Eric Schwitzgebel (2002; 2013). In **3.3** I will evaluate these two positions.

3.1 BELIEF and representationalism

On the representationalist account, to have a belief is to have an internal representation, what I am calling a *belief*, with a functional role that corresponds to the functional role of the concept of belief.

For example, for Celeste to believe that there are coffee beans in the cupboard is for her to have a representation with the content 'there are coffee beans in the cupboard' stored somewhere in her brain and for that representation to function as a belief in its interactions with input, other states and output.

Celeste's representation of coffee beans in the cupboard may be caused by her opening the cupboard and seeing a bag of beans in there, it may interact with other states to cause a variety of outputs such as retrieving the coffee beans in the cupboard, inferring that she has the materials required to brew a fresh cup of coffee, or deciding that she will save the coffee beans for a coffee the following morning when it tastes best.

There has been much debate amongst representationalists about the nature of *belief* representations. Fodor (e.g. 1975) argues that they are sentences in a language of thought, Camp (2007) argues they can be map like, while Dretske (1988) argues that they are representational systems that track features of the environment. The generalised account of the concept of belief is compatible with multiple *belief* formats. Provided they play the right sort of role, a single agent's complement of *beliefs* could be in a range of different formats. Indeed, the diversity of the concept of belief suggests that multiple representational formats are likely.

Most introductory books to the philosophy of mind will say that beliefs are propositional attitudes. A propositional attitude being, roughly, an attitude towards a proposition. This is often explained as the type of state that can precede a 'that' clause in everyday language. For example, Sylvie can believe that it will be sunny all day, and also she can *hope* that it will be sunny all day, *desire* that it be sunny all day etc. But we should be careful that this characterisation doesn't mislead us into thinking that belief is necessarily language involving. This focus on language results in a kind of linguistic bias, and results in language orientated accounts of the functional role of belief such as Schwitzgebel's (2011, see chapter 1).

Certainly, some beliefs can be attitudes towards linguistic propositions, but it isn't clear that all beliefs are propositional attitudes in this way. While mastery of language would seem necessary to believe something like 'Cassius and Brutus thought that assassinating Caesar would restore the Republic', it is less clear that more simplistic beliefs, such as 'there is food in the kitchen', require language. Indeed, the smell of frying sausages is apt to attract hungry humans and dogs alike.

While most would accept that language-using adult humans can have beliefs that are attitudes to linguistic propositions, we should not want to narrow their set of beliefs to just those that involve language. This would clearly exclude languageless animals and preverbal infants from having beliefs, but would also put unwanted constraints on the range of adult human beliefs. It seems clear that we also have beliefs that can have sensory, imagistic or map-like content. I can have imagistic beliefs about what a university building looks like, for example, but I may never have verbalised them before. When someone asks me for directions to that building, I will visualise the building before describing it to them, thereby translating the belief into linguistic format – "The Arts Tower? Yes, walk down this road and you can't miss it, it has several concrete legs supporting a tall bluish-greyish glazed tower. There is a kind of raised walkway going into it from an adjacent building."

Beliefs about taste are notoriously hard to verbalise. Describing how something tastes to someone else usually involves listing other things that taste similar. We can all think of the strange comparisons a TV wine expert may use to describe the taste of a wine they are reviewing. My vegetarian partner has previously asked me to explain what anchovies taste like, and why it is that some people love them, and some hate them. The best I can do is that "they are like a fishy caper", but that is of little use as she has no idea what a 'fishy' taste is.

Other candidates for non-linguistic beliefs could take the form of spatial representations. Your belief that there is one more stair could cause you to trip at the bottom of the staircase. Again, beliefs in this format need to be translated when we want to verbalise them. When a stranger asks for directions to a local landmark, we may need to walk the route in our head before giving the correct sequence of rights and lefts.

In addition to not being tied to any one format, the representations underlying the generalised concept of belief need not be in anyway localised within the brain. We need not take the popular 'belief box' metaphor of Nichols and Stich (2003) literally. The generalised account of BELIEF is compatible with fragmented belief storage (Quilty-Dunn & Mandelbaum, 2018, p.2358) in which "…our beliefs are stored in disparate, perhaps mutually inconsistent fragments.".

To have a belief, on the representationalist account, is to have a *belief* state that plays the right kind of functional role, stored someway and somewhere in the brain.

3.2 Dispositionalism

In contrast to representationalism, advocates of dispositionalism argue that to have a belief is the disposition to produce certain output when particular input conditions obtain. For Celeste to believe that there are coffee beans in the cupboard is for her to be disposed to act in ways as if the proposition 'there are coffee beans in the cupboard' is true. So, for example, when she has a hankering for coffee she may go to the cupboard to retrieve the beans, when asked by a guest if she has any coffee she would reply that she has some beans in the cupboard, and she would exhibit surprise if the cupboard were opened to reveal it to be empty.

In this section I will set out traditional dispositionalism as articulated by Ryle (1949) and the arguments against it, before moving on to the more recent form of dispositionalism put forward by Schwitzgebel (2002; 2010; 2013).

Dispositional properties are commonplace. Well-worn examples include the property of being soluble as the disposition to dissolve in water, and fragility as the disposition to break or shatter on impact. We can add to this list Ryle's (1963 [1949], p.43) example, evocative of life as a philosopher in the mid-twentieth century, of the property of being a smoker as "to be bound or likely to fill, light and draw on a pipe in such and such conditions."

To have a disposition is "to be bound or liable to be in a particular state, or to undergo a particular change, when a particular condition is realised." (Ryle, 1963 [1949], p.43)

More formally, Schwitzgebel (2002), p.250, states that:

> "Dispositions can be characterized by means of conditional statements of the form: If condition C holds, the object O will (or is likely to) enter (or remain in) state S. O's entering S we may call the manifestation of the disposition, C we may call condition of manifestation of the disposition, and the event of C's obtaining we may call the trigger."

Schwitzgebel's formulation fits simple properties such as solubility and fragility rather well. Their disposition consists of a single condition of manifestation and a single manifestation. Something soluble, such as a sugar cube, will dissolve when put in water. Something fragile will break on impact. All there is to something being soluble is that it has the disposition to dissolve when put in water[23]. Ryle called dispositions with a single condition of manifestation and a single manifestation, 'single-track dispositions'.

Ryle contrasts single-track dispositions with what he calls higher-grade dispositions, which are now more commonly referred to as multi-track dispositions. Ryle (1963 [1949], p.44) uses the work of Jane Austen to illustrate:

> When Jane Austen wished to show the specific kind of pride which characterized the heroine of Pride and Prejudice, she had to represent her actions, words, thoughts, and feelings in a thousand different situations. There is no one standard type of action or reaction such that Jane Austen could say 'My heroine's kind of pride was just the tendency to do this, whenever a situation of that sort arose.

[23] Or another liquid with a suitably high water content such as a cup of tea.

A multi-track disposition has multiple realising conditions and multiple manifestations. According to Ryle, beliefs are dispositions of this multi-track variety.

As I noted above, the belief that there are coffee beans in the cupboard is not simply to respond in the affirmative when asked if there are coffee beans in the cupboard. If the belief that there are coffee beans in the cupboard is a disposition, then there are potentially an indefinite number of realising conditions and an indefinite number of potential manifestations.

A particularly influential objection to the idea of belief as a multi-track disposition, which Schwitzgebel attributes originally to Chisholm (1957), I shall call the **argument from interactive function**. In short, the argument is that the output manifested by a particular belief is highly sensitive to an agent's other beliefs and desires. Belief as a multi-track disposition is therefore not enough to explain or predict an agent's output, we also need to understand how that belief interacts with input and with other internal states in producing output. If our aim is to reduce belief to observable behaviour, then one cannot invoke other beliefs when explaining observable behaviour without circularity.

Consider Celeste and her coffee beans. This evening Celeste is entertaining her friend Hans. After eating a so-so risotto for dinner, Hans asks Celeste if she has any coffee. Celeste believes that she has some coffee beans in the cupboard. However, she also believes that there are only enough beans to make a single cup of coffee and she very much wants to enjoy that single cup the following morning with her breakfast. Deciding that it is easier to avoid a potentially awkward conversation, Celeste tells Hans that she has run out of coffee.

The behaviour that would naturally manifest the belief that there are coffee beans in the cupboard, answering in the affirmative when questioned if she had any coffee, is overridden by Celeste's desire for a cup of coffee the following morning and her aversion to potentially awkward social situations.

If Celeste's surrounding beliefs had been different, then so would her behaviour. If rather than being short of coffee, she believed that she was fully stocked, then she would have answered Hans' question differently – she may have been happy to make him a coffee.

If we are going to effectively explain Celeste's behaviour when she lies to Hans about having coffee in the cupboard, then we need to attribute to her a mental state, we need to have information of how that mental state interacts with other mental states, and how this interaction gives rise to Celeste's behaviour.

The fact that the output produced by a belief is dependent on interactions with other mental states also raises a related objection, which I will call the **argument from absence of behaviour**. According to dispositionalism, for an agent to have the belief that P is for them to be disposed to behave as if P were true, so if an agent exhibits no manifestation of this disposition then they cannot be said to have the belief that P. However, it is possible to imagine situations in which we would say that an agent has the belief that P but does not behave as if that were the case.

The classic example of this (Schwitzgebel, 2002) is of an agent in a situation in which it would be potentially dangerous to reveal their beliefs. Think of the character Winston Smith in the novel *1984* (Orwell, 1987 [1949]). Smith lives under constant surveillance in a totalitarian state. While he believes that the state machinery is falsifying documents and that the regime should be overthrown, Smith knows that any outward manifestation of these beliefs would swiftly result in arrest and torture. His anti-establishment beliefs need to remain well hidden. In order to survive, he must maintain a perfect façade of conformity, even down to the detail of his facial expression. Despite the lack of any outward manifestation of belief, as the reader has access to Smith's thoughts and feelings, they would not hesitate to attribute such anti-establishment beliefs to him.[24] Belief, therefore, cannot be purely behavioural dispositions.

A new dispositionalism

The argument from interactive function and the argument from absence of behaviour were persuasive and dispositionalism fell out of favour. In recent years, Eric Schwitzgebel (2002; 2010; 2013) has worked to resurrect dispositionalism by developing a **phenomenal, dispositional** account of belief and of attitudes in general. Schwitzgebel maintains that a dispositionalist account can avoid the argument from interactive function, but only if it abandons the goal of reducing beliefs to external behaviour. Thus freed, the dispositionalist can appeal to other mental states when explaining behaviour and to allow phenomenal and cognitive dispositions in addition to the more familiar behavioural dispositions.

In addition to bolstering dispositionalism against objections, Schwitzgebel also offers a positive argument in favour of his phenomenal dispositional account. He argues that his account is much better equipped to explain what he calls **in-between** beliefs (Schwitzgebel 2002; 2003; 2013). Below I set out Schwitzgebel's account in some detail before discussing its strengths and weaknesses in section 3.3.

Central to Schwitzgebel's account is the concept of a **dispositional stereotype**. For Schwitzgebel, a stereotype is "a cluster of properties we are apt to associate with a thing, a class of things, or a property." (Schwitzgebel, 2002, p.250). A stereotypical cactus, for example, would have several properties, it would be a living plant, have a thick fleshy stem, no leaves, and a covering of sharp spines. While many plants will possess all of these properties and fit the cactus stereotype perfectly, it is not necessary to have all of the properties in order to be a cactus. Some cacti will be missing one or more of these properties, but they can still be considered a cactus. *Maihuenia patagonica,* for example, is an atypical cactus that is covered with spines and has a thick fleshy stem, but it does have leaf structures.

A dispositional property, as we established above, is the property of an object to enter or stay in a state when particular conditions are realised. Combining these two ideas then, we arrive at the concept of a dispositional stereotype – a stereotype that consists in a cluster of dispositional properties.

Personality traits provide a useful illustration. Being calm, for example, is just a matter of being disposed to react in certain ways in certain situations. Sylvie is a calm person. When a car drives through a puddle splashing her, she says "it's fine, I'll soon dry out". When the wrong food is brought out to her in a restaurant, she says: "this isn't what I ordered, but no problem. This looks nice too." When she realises that a piece of writing she has been working on has failed to save correctly, she smiles ruefully and starts again. There is nothing more to being calm than having a cluster of such dispositions.

According to Schwitzgebel, the same is true of beliefs: "To believe that P, on the view that I am proposing, is nothing more than to match to an appropriate degree and in appropriate respects the dispositional stereotype for believing that P." (Schwitzgebel, 2002, p.253)

But which dispositions? The list of dispositions associated even with simple beliefs, such as the belief that there are coffee beans in the cupboard, could be indefinitely long. Clearly for each belief we can't specify an exact list of all the potential dispositions. Rather, the belief that P consists in that cluster of dispositions "that we are *apt* to associate with the belief that P" (Schwitzgebel, 2002, p.251).

[24] Smith was unable to maintain his façade and ended up being tortured in room 101 and eventually shot, but the argument stands.

As with other stereotypes, the properties associated with a dispositional stereotype are not a list of necessary and sufficient conditions for membership of a category. Things can match a stereotype to a greater or lesser extent. While something that has all the stereotypical properties associated with a category would indisputably be considered a member of that category, other things that only match a portion of the properties may still be considered to be a member of a category, but not a stereotypical member. Sylvie may in all other respects be a calm person but is unable to keep her cool under very specific circumstances, say on witnessing cruelty to animals. Under such circumstances she will display anger, but despite this specific trigger most would still consider that Sylvie was, overall, a calm person.

It is not possible to define what matching a stereotype to an "appropriate degree" is. Different individuals will draw lines in different places and weight properties in different ways. I have a friend that once insisted that for something to be a pie, the filling must be completely enclosed in pastry. He vehemently maintained this position against a majority who were in agreement that something which consisted of filling in a dish with a pastry top was a type of pie, albeit an atypical one.

While the idea of a dispositional stereotype arguably has some advantages when considering cases of in-between belief, something that I will discuss in the next section, it does not in itself do anything to avoid the argument from interactive function. This can only be done by abandoning what Schwitzgebel sees as one of the main motivations for the original dispositionalist accounts – advancing the materialist account of the mind. Dispositionalism had provided a way of reducing beliefs and other mental states to observable behaviour, thereby providing an acceptably materialist account of something that had previously proved extremely difficult to explain with just the resources of the physical world.

According to Schwitzgebel, the requirement to reduce belief to the observable is what makes dispositionalism vulnerable to the argument from interactive function. Once the dispositionalist abandons this requirement, they are free to include dispositional manifestations other than observable behaviour in a *belief's* dispositional stereotype. They can also accept that an agent can have a belief but not match the dispositional stereotype due to that agent's other beliefs and desires.

Phenomenal and cognitive dispositions

I noted above that one of the consequences of the argument from interactive function is that the influence of other mental states can mean that the connection between a belief and observable output may be very weak. Winston Smith's anti-establishment beliefs, for example, could only manifest observably on pain of torture and death. Smith's beliefs then, could not be reduced to behavioural dispositions because there were no behavioural manifestations of his belief.

However, by dropping the requirement that a belief must be reduced to the observable, the dispositionalist is free to add dispositions that do not have observable manifestations to their stereotypes[25]. They are now free to include dispositions that have phenomenal or cognitive manifestations[26].

Going back to the example of Celeste and her coffee, her desire to hide the fact that she has coffee in the cupboard from her guest means that she takes care to ensure that there are no observable

[25] Arguably, that arch-dispositionalist Ryle also allowed that the dispositions that characterise a belief could include such things as "imaginings" (Ryle, 1963 [1949], p44) and he took great care to demonstrate that the secrecy of things that happen "in one's head" were "not the secrecy ascribed to the postulated episodes of the ghostly shadow-world. It is merely the convenient privacy which characterizes the tunes that run in my head and the things that I see in my mind's eye." (Ryle, 1963 [1949], p35)

[26] Schwitzgebel (2013 p19) describes accounts that admit observable behaviour only as **behaviourist** and those that admit phenomenal and cognitive phenomena as **intellectualist**. His account, which admits both, he describes as **liberal dispositionalism**.

manifestations of her belief that there is coffee in the cupboard. Internally though, it is a different story. When, Hans asks if there is any coffee, Celeste responds that there is none, but at the same time Celeste may picture the jar containing the single serving of coffee; she might say to herself “I have got coffee, but just enough for one cup”; she might even picture herself enjoying that cup of coffee in the early light of the following morning. These we would call phenomenal dispositions, that is, dispositions that have a phenomenally conscious manifestation. We can also imagine dispositions that have a cognitive manifestation, such as an inference, decision or judgement. In our example, Celeste’s belief that there is coffee in the cupboard, in combination with her desire to have that coffee the following morning, leads to an inference that if she were to serve up that coffee there and then, she would not be able to have coffee the following morning. She judges that Hans may think her a little selfish if she tells him that, yes, she does have coffee, but she isn’t going to make one for him. Celeste therefore decides that she is going to conceal the fact that she has coffee from her guest Hans.

Schwitzgebel argues that allowing that dispositions can be phenomenal and cognitive as well as behavioural enables the dispositionalist to meet the argument from absence of behaviour. While Celeste doesn’t exhibit the behavioural dispositions that we would be apt to associate with the belief that there is coffee in the cupboard, her phenomenal and cognitive dispositions are enough for a dispositionalist to attribute the belief to her.

***Ceteris paribus* dispositions and interaction with other mental states**

The argument from interactive function holds that the output that beliefs are apt to produce is highly sensitive to an agent’s surrounding beliefs and desires. This has the consequence that dispositional manifestations that we would associate with a particular belief may not occur when the trigger for that disposition obtains. For example, taking an umbrella on a walk to the shops is something that we would consider to be an appropriate manifestation of the belief that it is raining. Josh, however, believes that it is raining but does not take an umbrella with him on his walk to the shops, even when offered one, because he also believes that carrying an umbrella is something only middle-class people do, and he desires not to appear middle-class.

Schwitzgebel accepts that dispositional stereotypes only hold *ceteris paribus*. However, this raises a potential issue with his account. If to have a belief is to match a dispositional stereotype, where there are deviations from that stereotype, how can we distinguish between cases of deviation due to an absence of belief and cases where the agent has a belief but circumstances prevent it from manifesting? The latter are what Schwitzgebel calls “excused” deviations from a stereotype (Schwitzgebel, 2002, p.254).

The first thing to note is that once we allow phenomenal and cognitive dispositions into our stereotypes, then it is quite difficult to think of examples where an agent can completely deviate from a dispositional stereotype and yet we would still intuitively think they have the belief.

In the example of Josh his deviation is minor, and he may still reasonably accurately match the dispositional stereotype for the belief that it is raining. On looking outside, he may tut and say: “look, it’s raining”. And while he may scoff at the offer of an umbrella, he still resolutely puts up his hood before stepping outside.

With our other examples it is more difficult. Celeste and Winston take care to ensure that there are no external manifestations of their beliefs. They still, however, will to some extent match the dispositional stereotype if we accept phenomenal and cognitive dispositions into our dispositional stereotypes. If Winston Smith was not only outwardly compliant but also internally affirmed his love for the ruling regime then it would be natural to think that he had no anti-establishment beliefs.

Perhaps though, we can still imagine examples in which we would naturally think that an agent has a belief but fails to match the dispositional stereotype. Say that at the moment that Hans asks if she has any coffee, Celeste is momentarily distracted by something that has been weighing on her mind recently, she automatically replies "No". Her thoughts are taken up by the matter that she is distracted by and not visualisations of the coffee in the cupboard, or any other phenomenal or cognitive manifestations.

According to Schwitzgebel, the process of determining whether a deviation is excusable is the same for dispositional stereotypes as it is for scientific laws. There are two broad considerations. Firstly, we would need to discover the range of circumstances in which we can we expect deviation, in order to determine how general the stereotype is. If only a small proportion of agents with the belief that P match the stereotype, then it is a poor generalisation with little practical use. Secondly, how important is the deviation to the generalisation?

An example of an unimportant deviation would be Josh not taking an umbrella despite believing that it is raining. We might think that carrying an umbrella would seem to be on the periphery of the stereotype for believing that it is raining. We wouldn't think that Josh believes that it is raining any less than someone that does accept the offer of an umbrella. Nor would we think that Winston Smith believed that the government should be overthrown any less than someone who was not as able to disguise their behaviour as well as he is. There are clear excusing conditions in both examples. We might be less inclined to excuse Celeste's deviation when she is distracted. At that moment, we might think that she didn't believe that she had coffee in the cupboard. Or, weighing up the two factors, we might consider that while her deviation is important, her behaviour over a longer time period demonstrated that even during that moment she did in fact believe that there was coffee in the cupboard, and the circumstances excused the deviation.

Schwitzgebel concedes that there is no clear principle that will allow us to choose between what is an excusable deviation from a dispositional stereotype and what is not, but he does not see this as an issue with his account. We do not see an issue with scientific laws holding *ceteris paribus*. Nor do we have a problem with the fact that the excusing conditions for deviations from scientific laws cannot be detailed exhaustively. Why then, would this be considered an issue for dispositional stereotypes?

By allowing that the cluster of dispositions associated with a belief can include phenomenal and cognitive dispositions as well as behavioural ones, and by accepting that other mental states can impact on a belief's dispositions, Schwitzgebel argues that his account is secured against the objections that were levelled against traditional dispositionalism. So far, I have set out Schwitzgebel's position as sympathetically as possible. In the next section, however, I will critically evaluate the dispositionalist account of belief and determine whether it provides a better account than representationalism.

3.3 Evaluating the positions

In this section I evaluate the two positions I have considered in this chapter. Firstly, in **3.31**, I will show the explanatory and predictive advantages of the representationalist account. In **3.32** I will then determine if there are any advantages of the dispositionalist account over the representationalist position. I will consider two potential advantages proposed by Schwitzgebel, firstly that dispositionalism matches well with our intuitions and secondly that there are cases of 'in-between' beliefs that cannot be explained with the resources of representationalism. I will show, however, that neither of these favour the dispositionalist account over the representationalist.

3.31 The advantages of 'deep' accounts

It is important to note the limited aims of the dispositionalist account. It is a theory of what it is for an agent to have a belief. The theory claims that we are justified in attributing a belief to an agent when that agent exhibits a pattern of output that we are apt to associate with that belief. Schwitzgebel calls this type of account **superficial** and contrasts it with what he calls **deep** accounts which aim at a functional description of the underlying structures that cause the characteristic pattern of output.

In this section I argue that deep accounts, such as representationalism, can provide far richer and more flexible explanations and predictions of output.

Schwitzgebel compares his approach to belief and other attitudes with personality traits. To have a personality trait is to act in a certain way in certain circumstances. Recall the earlier example of calm Sylvie, who manages to keep her cool in a whole range of testing circumstances. Sylvie matches well our dispositional stereotype for being a calm person. Personality traits can be useful ways of forming expectations about people's behaviour.

Sylvie, who we know as a calm person, leans her bicycle against the wall while she talks to friends. When a passer-by clumsily bumps into the bicycle and knocks it over, knowing that Sylvie is a calm person, we can expect her to react in a cool way. In contrast to her sister Sylvie, Celeste is known as a fiery character. If someone were to clumsily knock her bike to the ground, then we would expect an angry reaction. Of course, we don't need to know someone in order to form expectations based on a personality type. If we had never met Celeste but someone had told us that she was fiery, then we could make a good guess at the type of reaction we should expect if someone were to clumsily damage her bike.

To be effective, a personality trait needs to accurately describe a cluster of dispositions that tend to co-occur in individuals. To be calm, fiery, extroverted, narcissistic, etc. is just to match the dispositional stereotype apt for each of those personality traits. The form of prediction or explanation which we can make on the basis of a personality trait is a "subsumption under a generalisation" (Carruthers, 2013b, p.152). We predict that Celeste will react angrily to her bike being damaged because that is in general how fiery people react when they feel wronged.

Representationalist explanations citing belief are of a different kind, however. Carruthers (2013b in an example adapted from Harman, 1965) gives the example of the detective Poirot. Poirot was questioning a butler who is suspected of murder. The butler claims that he could not have committed the murder, as on the night of the murder he was not at the scene but rather had spent the night in a hotel in the next town over. Poirot, who knows that the hotel in the next town over is closed for the winter, is able to conclude that the butler is lying.

We don't explain or predict Poirot's conclusion because that is the type of conclusion that Poirot generally draws, or the type of conclusion that moustachioed Belgian detectives generally draw. Rather, we are citing his belief that the hotel was closed for the winter as a reason for his conclusion that the butler is lying. Poirot sees an inconsistency in his belief that the hotel in the next town over is closed and the butler's claim that he stayed there on the night of the murder. Both his belief and the butler's claim cannot be true. As Poirot is certain in his belief that the hotel in the next town over is closed for the season[27] he concludes that the butler is lying.

While the dispositionalist says that an agent has a belief because they exhibit a particular pattern of output, the representationalist account can provide an explanation as to why the agent displays that pattern. In terms of explanation and prediction, the representationalist account is far ahead.

The meagreness of the dispositionalist explanation can be seen when we consider cases such as that of Winston Smith. On the dispositionalist account, Smith partially fits the dispositional stereotype for having the belief that the government should be overthrown. That is to say, he has the phenomenal and cognitive, though not the behavioural, dispositions which we would be apt to associate with the belief that the government should be overthrown.

[27] As a very thorough detective Poirot has already journeyed to the next town over to confirm that the hotel is indeed closed and by questioning locals has confirmed that it has been closed since the start of the winter.

But why does Smith only partially match the dispositional stereotype? Why does he not display the behavioural output we would associate with anti-establishment beliefs? Dispositionalism can offer no insight. With the resources of representationalism, however, we can offer a rich explanation for the lack of behavioural output consistent with his belief – Smith believes that if the party learns of his belief that the government should be overthrown they will torture and kill him. Smith does not desire to be tortured and killed, so tries to hide his belief with outwardly compliant behaviour.

For Schwitzgebel, the cluster of dispositions that are associated with the belief that P, are those dispositions "that we are *apt* to associate with the belief that P" (Schwitzgebel, 2002, p.251). But the range of our dispositions that we are apt to associate with any given belief is going to be very long and very general. With the resources of representationalism, we can determine a specific narrow range of output that would be apt to a given set of beliefs, desires, and circumstances. We can form specific expectations and predictions of output and we can give detailed, causal explanations. Having this type of explanation of why Smith has to keep his beliefs to himself is key to understanding the novel.

Our explanation as to why Smith keeps his beliefs to himself is not due to the fact that we know that people that have those beliefs, those desires, and who are in that political situation have behaved like that in the past as we would on the dispositionalist account. We can do so because we can exploit a body of information that describes how mental states such as belief interact with input and each other to produce output. The process of determining what output is apt to associate with specific beliefs, something skipped over in Schwitzgebel's dispositionalist account, involves a process similar to that described by the deep account.

Schwitzgebel argued that by dropping the requirement to reduce *belief* to the observable, that dispositionalism was able to avoid the argument from interactive function. But acknowledging that other *beliefs* and *desires* may cause deviation from a stereotype does nothing to allow the dispositionalist account to explain deviations from stereotypes. Representationalism, however, has the resources to explain in detail why any particular *belief* had only a partial impact or no impact at all on the behaviour of the agent.

Because of the explanatory and predictive benefits of the representationalist account, all things being equal it should be preferred. We should only want to embrace the more superficial dispositional account if there were fatal objections to the representationalist account or if the dispositionalist account were able to provide an account of phenomena that could not be explained with the resources of representationalism. I assess whether this is the case below.

3.32 Are there advantages to the dispositionalist account?

I showed in the previous subsection how deep theories of belief such as representationalism can provide far richer and more flexible explanations and predictions of output. So, for what reasons could we be convinced to accept the dispositionalist position over the representationalist? Schwitzgebel suggests two such reasons. In this subsection I will evaluate and reject both of them.

Alien attitudes

One proposed advantage that the dispositional account is claimed to have is that it matches well with certain folk intuitions. Consider what Schwitzgebel calls **alien attitudes** (2013). Imagine a situation in which we are visited by an alien called Watt[28]. Watt looks like a human in all respects with the exception that his ears are the opposite way around to a human's. When out in public on a trip to the shops, Watt wears a deer stalker hat which covers his ears and he manages to easily pass as a regular human. To explain Watt's behaviour, it would seem natural to say that he **desires** to remain

[28] I have borrowed this character from the 1990s BBC children's series 'Watt on Earth'

incognito, and that he **believes** that the deer stalker hat will cover his ears and allow him to pass as human. However, despite looking very similar to a human, Watt's behaviour has very different causes to ours. He does not have a brain as such, nor any other cognitive organ. The dispositions that we interpret as indicative of a belief have no categorical basis. We might imagine that, like Schwitzgebel's Beta Hydrian aliens (2013, p.84), beneath his human like exterior, Watt is made entirely of undifferentiated balsa wood.

While this is something that we might be able to imagine visually, it is more difficult to imagine how such complex output could be produced in reality. Setting this major obstacle to one side, the example is intended to show that if the representationalist insists that belief is possession of a particular mental state, then Watt clearly doesn't have this mental state. Yet, ordinary folk would happily attribute to him the belief that his deer stalker hat will hide his ears. The example of alien belief shows that where the possession criteria of representationalism and dispositionalism diverge, the folk favour dispositions.

But we can think of other examples that pull our intuition in the other direction. A less far-fetched example than aliens made of balsa wood can be found in Botterill and Carruthers (1999, p.32) – that of a chess computer that uses a brute-force algorithm. The algorithm sorts through all the potential moves that it could make, then all the potential moves that could be made in response to that move, then all the potential moves that it could make in response to the response and so on. After a certain amount of time or a certain number of calculations, the computer selects the move that leads to the best outcome, for as far ahead as it has had chance to calculate.

If you were playing chess against a human opponent it would be natural to think, for example, that they believed that a certain move may bolster their king's defence. However, if the brute-force chess computer played that same move, then we would be less inclined to attribute that belief to the computer, because of our knowledge of how the move was decided upon.

The chess computer's human opponent may find it useful when trying to anticipate their opponents moves to attribute to them certain beliefs, desires, and intentions. However, this would be no more than a useful way of thinking, not to be taken literally. Considering our knowledge of how the chess computer is calculating the moves, it would not be appropriate to attribute these beliefs to the chess computer. Similarly, if we found that the alien, Watt, was simply using a brute-force algorithm to calculate canned social responses, then it wouldn't seem appropriate to attribute beliefs to him.

Contrary to Schwitzgebel's claims, when it comes to belief, it is not just *what* we do that counts, but *how* we do it. Cognitive architecture matters. Rather than constituting a belief, dispositional manifestations[29] provide evidence that an agent has a particular internal structure.

Of course, when it comes to attributing beliefs, the folk use observable behaviour. Very rarely would we have any way of determining whether any deep conditions, such as possession of a particular representation, obtain. But in these examples where we do have the knowledge that the observable behaviour is not caused by the interaction of states matching the functional description of belief, desire, intention and the like, it intuitively feels inappropriate to attribute those states.

As shown above, intuitions can pull in both directions and do not conclusively favour the dispositional or the representationalist account. But should our intuitions have any bearing on which account we choose? As we established in chapter 1, the eliminativist challenge that we are evaluating is not a threat to everyday folk talk, but usage of the concept of belief in cognitive science. We are not trying to determine which account of belief matches best with our everyday intuitions, but which account would

[29] In practice these dispositional manifestations would almost always be observable behaviour, but they could also be phenomenological and cognitive if we had appropriate access. If we have the privileged position of reading about a character in a novel, for example, or if we were attributing beliefs to ourselves and something like Carruthers' Interpretative Sensory-Access theory of self-knowledge is correct (Carruthers, 2011)

be most suitable to integrate into science as a starting point for theorising about cognitive architecture. Our folk intuitions should have limited bearing on which account is favourable for this purpose. Rather, the considerations of explanatory and predictive power as discussed in the previous chapter should take precedence. Folk intuition should be, at best, a minor factor in determining which account to prefer and, as we have established in this section, a factor that favours neither account.

In-between belief

Schwitzgebel's primary argument in favour of his account is that dispositionalism is better able to explain instances of what he calls in-between believing (Schwitzgebel, 2002; 2003; 2013). In-between beliefs are those cases in which it isn't clear whether someone has the belief or not. On the dispositional account, having a belief is a matter of matching to a greater or lesser extent one or more dispositional stereotype, there is no further fact of the matter. In contrast, Schwitzgebel argues, if whether someone has a belief or not is determined by possession of a certain internal structure, then we are forced to say that the agents described in the cases of in-between believing either have the belief or they don't.

Examples of in-between beliefs include:

1) Juliet, a white, politically liberal professor who rails against conventional standards of beauty and explicitly states her belief that all races are equally beautiful, but responds more favourably to images of conventional white beauty (Schwitzgebel, 2013)
2) Antonio, who when in church listening to a sermon on the magnificence of creation, feels sure that there must some kind of higher power, yet at other times views religion as metaphorical and does not defend religious belief during debates (Schwitzgebel, 2003)
3) Ben, who receives an email to tell him that a bridge on his daily commute will be closed, but the following day still takes his usual route (Schwitzgebel, 2010).

In all these examples it is tempting to ask, what do they really believe? Does Juliet really believe that all races are equally beautiful, does Antonio really believe that God exists, does Ben really believe that the bridge is out? But in all three cases, there doesn't seem to be a clear answer. To say that each of the agents above must believe one way or the other seems counterintuitive, it would oversimplify each case. The most natural response would seem to be that in some respects they believe that A, while in others they believe that B[30].

Schwitzgebel argues that the dispositionalist account can comfortably handle the uncertainty of cases such as these, reflecting our intuitions. Ben, for example, matches to some extent the dispositional stereotype for believing that the bridge is closed. He may say to his wife on receiving the email about the bridge "Arrgh! The bridge over the river is down. What a nightmare!". In his head he may visualise the closed bridge, he may start to work out a new route, and he may calculate the additional time it will take to travel to work. All of these seem like manifestations of a dispositional stereotype appropriate for an agent that has a belief that the bridge is closed. However, Ben sets off for work in the morning as usual, taking his usual route – a behavioural manifestation consistent with the belief that the bridge is open.

While Schwitzgebel sees his account as "a description of the conditions under which particular beliefs can properly be attributed to human beings" (Schwitzgebel, 2002, p.258), in this instance we would want to hold off making an attribution and remain agnostic, leaving it as Ben having a strong match for the dispositional stereotype appropriate for believing that the bridge is closed and also matching to some extent the dispositional stereotype appropriate for believing the bridge is open. The

[30] As mentioned at the beginning of the chapter, a dispositionalist could use Schwitzgebel's treatment of in-between beliefs to defend the concept of belief against Jenson's (2016) charge of fragility that we discussed in chapter **1**. It is not that some methods of observation are detecting belief and others aren't, it is just that individuals in-between believe and this is picked up by some methods of detection and not others. This isn't a line of argument I will develop as I endorse the representational account.

dispositionalist is not forced to make a choice one way or the other. “Once all the relevant dispositions have been made clear, the case is closed. There are no further facts to report” (Schwitzgebel, 2002, p.262).

We can give a similar treatment to the other examples. In some respects, they match the dispositional stereotype for belief A, while in other respects they match the dispositional stereotype for belief B. This approach provides a nice match to our intuitive thoughts about these cases.

Representational accounts of belief however, according to Schwitzgebel, are unable to do justice to our intuitions in these cases. If belief possession is determined by the presence of a specific internal representational state, then an agent either has that state or they don’t. Schwitzgebel draws on the popular metaphor of the belief box, which has been used extensively by Fodor (1987) and Nichols and Stich (2003), amongst others. In this metaphor, to have a belief is to have a representation in the belief box, which is to say that it interacts with other representations in a characteristically belief like way. The metaphor encourages us to think that there is either a belief in the belief box or there isn’t. Juliet either believes that all races are equally beautiful, or she doesn’t.

Appealing to degrees of confidence in a belief does not help the representationalist explain the cases of in-between belief. It isn’t that Juliet is only 70% confident that all races are equally beautiful, or that Ben is 10% confident that the bridge will be open and decides to take a punt.

Unlike the dispositionalist then, the representationalist does not have the resources to properly account for our intuitions about these cases. Or so maintains Schwitzgebel. This characterisation of representationalism, however, is somewhat unfair. I believe it is indeed possible to provide a satisfactory representationalist account of cases of in-between believing.

The criticism that representationalism is unable to handle cases of in-between believing rests on a failure to recognise the key distinction between “failures of storage and failures of access” (Quilty-Dunn & Mandelbaum, 2018, p.2359). From the fact that a belief has not played a role in the production of a specific output, it does not follow that that belief is not present, it could simply mean that whatever process led to said output was unable to access the representation during its working. There are several reasons why this would be the case. Cognitive load could prevent the belief from being accessed; or some beliefs may only be accessible to all systems under certain circumstances, while some beliefs may only ever be accessible to modular systems.

Taking the example of forgetful Ben, a representational explanation of his behaviour might proceed something like this. After reading the email informing him that the bridge is out, Ben forms the belief that the bridge is out. You might say that he stores this in his belief box, if you are fond of that metaphor. This belief then plays a role in production of several pieces of his output – his exclamation to his wife, his visualisation of the bridge being out, and his planning of an alternative route to work. The following morning, he wakes up at 0645, showers, eats his toast, drinks his espresso, gets on his bike, and turns right at the end of the road – straight towards the closed bridge.

Ben’s belief that the bridge is out is not accessed during his whole morning routine and therefore cannot have an influence on his behaviour. One reason for this failure of access may be the fact that Ben goes through exactly the same morning routine five days a week and it has become largely automatic. Further factors contributing to the failure of access may be Ben’s tiredness after a restless sleep, or the fact that his mind is occupied with thoughts about his marital problems.

While Ben does have the belief that the bridge is out, his actions are straightforwardly explained as a failure of access. What we might call a bout of temporary forgetfulness and something I imagine we are all familiar with. The representationalist can, therefore, account for the in-between case of forgetful Ben.

We could also construct similar explanations for the case of Juliet. Her beliefs that all races are equally beautiful is fully consciously accessible, meaning that this belief plays a role in Juliet's speech and consciously planned behaviour. Her belief that white people are more attractive however, is not accessible to the same extent as her egalitarian belief. It may be inferentially insulated to some extent, either permanently or due to certain cognitive conditions. Arguably, as pointed out by Carruthers (2013b), this latter state should not properly be called a belief and should rather be characterised as something like a preference or an *alief* (Gendler, 2008a & 2008b). I will return to this potential reclassification of these states in the final chapter.

The above discussion shows that cases of in-between belief can be adequately accounted for with the resources of representationalism. Cases of in-between believing do not, therefore, favour the dispositionalist over the representationalist account.

In this chapter I have demonstrated that the representationalist account of belief possession is to be preferred to the dispositionalist. Schwitzgebel's attempt to resurrect dispositionalism fails to meet the argument from interactive function and the argument from the absence of behaviour. While Schwitzgebel's modified dispositionalist account allows him to accept that other states can cause deviations from dispositional stereotypes, it has no resources to explain such deviations. In Orwell's *1984*, that Winston Smith only partially meets the stereotype for having anti-establishment beliefs is at best an unsatisfactory and incomplete explanation of the reasons for his compliance with the establishment. The representationalist account has a clear advantage over the dispositionalist in terms of the richness and flexibility of the explanations and predictions of human output that it can provide.

I looked at two potential reasons for preferring the dispositional account. Firstly, that it better matches with our intuitions in the cases of alien attitudes. This was rejected because we can also generate intuitions in the other direction and the usefulness of using intuitions for determining belief possession is questionable. I next looked at cases of in-between belief, which Schwitzgebel argues are problematic for the representationalist account. However, I showed that these cases can be adequately explained with the resources of representationalism. With the clear advantages of the representationalist account and no positive reason to prefer dispositionalism, we should reject the dispositionalist account. To have a belief is to have a representational state that plays the functional role of a belief in our cognition.

The representationalist account is preferable because of the explanatory advantages, but it is perhaps more 'risky' in terms of the potential for elimination. However, the susceptibility of the representationalist account to the eliminativist challenge is precisely its strength. An account of belief that cannot adapt to take account of advances in cognitive science is of no use to scientists. The representationalist account provides the starting point from which cognitive science can investigate the boundaries of belief, determining which types of informational states genuinely play the role of belief and which, if any, play distinct roles.

4. BELIEF in Cognitive Science

In chapter 1 I identified that the key factor determining whether the concept of belief should be pragmatically eliminated from scientific discourse was whether the concept has scientific utility. Also in chapter 1, we drew a lesson from the case study of innateness – that elimination is domain specific. Cognitive science consists of many disciplines. Concepts that have no utility in one discipline may have utility in others. To show that BELIEF has scientific utility, we need to show that it has utility in one or more domain.

In this chapter I will describe a case study of a scientific debate in which the generalised concept of belief has played a key role, or so I will argue in chapter **5**. I will focus on the debate within developmental psychology between **rich** and **lean** explanations of experimental findings demonstrating mental understanding abilities in preverbal infants. The primary question in this debate is whether infant subjects achieve their mental understanding abilities by attributing a representation to the agent[31].

One popular explanation of children's mental understanding abilities is that they are employing a **theory of mind**. Employing a theory of mind is to implicitly assume that a variety of different types of mental state interact with input and each other to produce an agent's output. A theory of mind is a conceptual framework that models this system. A subject can use these conceptual resources to explain, predict or form expectations about a target agent's output by attributing to them an initial mental state or states, and modelling the causal steps that will lead/have led to the production of that agent's output. We can see that belief-desire psychology, which I discussed in chapter 2, is a theory of mind[32].

How do I know that Watt, the alien with back to front ears, will wear a hat when he goes to the shops? Because I can use my concepts of BELIEF and DESIRE to model the interactions of his internal states of *belief* and *desire*. I attribute to Watt the desire to remain incognito and the belief that a hat will cover his ears and make him look human. On the basis of these attributions I can then predict that Watt will wear the hat to achieve his desired state.

As a theory of mind consists of representations of representational states, mental understanding abilities that exploit a theory of mind can be referred to as **metarepresentational**.

Rich theorists argue that the mental understanding abilities demonstrated by preverbal infants are achieved by their using a theory of mind (e.g. Southgate, 2013). That is to say, they attribute metarepresentations to agents and form expectations of an agent's behaviour on the basis of these attributions. Lean theorists (e.g. Perner, 2010) argue against the metarepresentational explanation and instead claim that the infant subjects are exploiting a variety of behavioural rules that directly link observed behaviours with expected behaviours.

I will set out the case study in some detail in this chapter, showing how each account explains the experimental data. I will not attempt to develop a conclusive argument in favour of the rich account but

[31] Where I refer to the 'subject' I mean the individual using their mental understanding abilities to explain or predict output. Where I refer to the 'agent' I mean the individual whose output is being explained/predicted.
[32] Not everybody accepts this 'theory-theory' account of mental understanding abilities. The major alternative view is *simulationism*. According to this account, mental understanding abilities are achieved by simulating others' minds using out own mind. While I won't go into detail on the theory-theory vs simulationism debate, it is worth noting that a pure simulation account cannot account for systematic errors in mental state reasoning, and that theory-theorists can accept that simulationism is used as a device to supplement theory-theory reasoning (Saxe, 2005). As we shall see, the consensus view amongst developmental psychologists is that children employ a theory of mind.

do note that the rich account is to be favoured over the lean. I will show that the rich account has not only been used to explain empirical data, but has also inspired a fruitful research project and a series of successful experiments designed to rule out the rival lean theory. Before setting out the case study I will describe how concepts can play *direct* or *indirect* roles in scientific theories, and show how general concepts can play a role in some scientific disciplines even where those general concepts have been previously fractured into several distinct sub-concepts and have become redundant in other disciplines.

In chapter **5**, I will move on to argue that the rich account of preverbal infant cognition directly involves the generalised concept of belief.

4.1 Direct and indirect concepts

One way in which BELIEF plays a role in the theories of developmental psychologists is as the state that infants attribute to agents to understand their behaviour. For example, we form an expectation that Celeste will look in the cupboard for the coffee beans because we attribute to her a belief that the coffee beans are in there. For the rich theorist, infants exploit the same mechanism. For example, in a very influential paper Onishi and Baillargeon (2005, p.257) argue in favour of the rich interpretation of their data by stating that:

> "To explain these and the present results, it is more parsimonious to assume that infants attribute to others beliefs that can be shaped and updated by multiple sources of information than to assume that infants form an extensive series of superficial expectations linking different perceptions to different actions."

Other theorists are more cautious with their use of the term 'belief', Southgate and Vernetti (2014, p.8) write "…it is worth noting that what we demonstrate in this paper and describe as 'belief-based action prediction' does not necessarily imply that infants operate with a concept of belief as it is traditionally characterized.". While still others, such as Apperly and Butterfill, argue that infants' early success on false belief tasks is best explained by the attribution not of beliefs, but rather of belief-like states called 'registrations' which "…serve as proxies for beliefs both true and false: In a limited but useful range of situations, what someone believes about objects will match what they register." (Apperly & Butterfill, 2009, p.963).

In rich theories of infant mental understanding abilities, the infants are arguably attributing a BELIEF to an agent and forming an expectation of behaviour on the basis of that attribution. Here, BELIEF is clearly playing a role in a scientific theory. But this is only to show that BELIEF has a very limited and specific kind of scientific utility. In this case, BELIEF is only playing what I will call an **indirect** role in the theory. I will unpack this below.

In the social sciences, a concept can be said to play an indirect role in a theory when that concept is being deployed by the experimental subjects themselves. Where a concept appears in a theory indirectly, we need not endorse the usage or accuracy of that concept. The concept is rather an object or target of the scientific theory, rather than a direct part of the theory itself.

For example, consider the following hypothetical situation, an anthropologist finds that an isolated people predict and explain behaviour, not by invoking an agent's mental states, but by reference to interactions between various spirits. As an explanation as to why someone goes to the river, they say that the water spirit is in the ascendancy for that individual. As an explanation of why they sacrifice chickens, the tribespeople say sacrificing chickens pleases the hunting spirit and a happy hunting spirit makes the tribe's hunters act more bravely.

In her report, our anthropologist explains why a tribesperson sacrifices chickens "according to this tribe's way of thinking, how much bravery a hunter shows on a hunt is determined by how pleased the hunting spirit is. If the hunting spirit is happy, then a hunter will not back down, even when faced with an extremely fierce animal. If the hunting spirit is displeased however, the hunter is likely to flee in the same situation. The tribe believe that sacrificing chickens will please the hunting spirit and so they sacrifice chickens to ensure the bravery of their hunters."

In this example, while the anthropological theory of why a tribesperson sacrifices chickens involves the supernatural concept of the hunting spirit, we shouldn't conclude from this fact that the anthropologist endorses the concept of the hunting spirit. We wouldn't say that the hunting spirit is a useful anthropological concept. The concept features in the anthropologist's explanation only indirectly, as a concept that is used by the subjects of the explanation.

In our example here, note that the concept of belief *is* playing a **direct** role in the anthropologist's explanation. The practice of sacrifice is not explained by the hunting spirit, which may or may not exist, but rather the tribespeople's BELIEF in the hunting spirit. The hunting spirit concept is playing an indirect role, while BELIEF is playing a direct role.

When considering the case study of preverbal infants' mental understanding abilities in this chapter, it will not show that BELIEF has scientific utility if the concept only plays an indirect role in the theories explaining the experimental data. It is necessary to show that belief plays a **direct** role in the theories of cognitive scientists. What I am concerned to show then is not that the concept of belief is part of the folk framework that humans exploit to understand the minds of others, but rather that BELIEF is directly involved in the scientific theory that developmental psychologists have developed to understand that everyday folk framework.

I will argue that the developmental psychologists' claim that subjects attribute and ascribe metarepresentations is equivalent to claiming that subjects have BELIEFS about the representational states of the agents that are the targets of their mental understanding abilities.

In the case study, we will see that rich theorists argue that the mental understanding abilities that infant subjects display are based on their attribution of a false belief (or belief-like state) to the target agent. If my argument is correct, these metarepresentations could be accurately labelled as the subject's beliefs about the beliefs of the agent.

While technical terms such as attributing, ascribing and metarepresenting have replaced the label 'belief', I will show that BELIEF, the generalised concept developed by philosophers as described in section **2.3**, is still an important concept with a direct role in the theories of developmental psychologists. Indeed, the relabelling of BELIEF with new technical terminology is a clear demonstration that the concept has been successfully integrated into the theoretical frameworks of cognitive science.

4.2 General concepts in science

My aims are modest. I will show that BELIEF can and does play a useful role in at least one area of cognitive science. This is of course not to say that the concept still has a role to play in all areas. Indeed, in some areas of cognitive science, the broad brush of a generalised concept such as BELIEF may not be precise enough and scientists will look to draw distinctions between different types of belief-like states (although as I show in chapter 5 this may not be straightforward). But there are still other areas of cognitive science, in which the theoretical action is focused elsewhere, where a generalised BELIEF is sufficient to construct a theoretical model to explain the experimental evidence.

A similar example of the use of generalised concepts that still have value in science is the concept of **food** in behavioural ecology. Food, like BELIEF, is a very general concept. For an omnivore such as the bear, the category of food could consist of such diverse things as nuts[33], berries, insects, fish and other

[33] I am talking here of nuts in the culinary sense

mammals. We also know that food is not a simple single resource, but rather is a mix of many different macro and micronutrients that meet the various nutritional and energy needs of an organism. Many animals may need to vary their diet to ensure that it meets all their nutritional needs, and food intake may vary in the course of the seasons due to differing availability.

Despite the complexity within the generalised concept of food, we still see it playing an important role in some areas of science. A recent example is Beauchamp (2017), who proposes a new model of anti-predator vigilance in prey species. Predator vigilance, essentially the amount of time that an individual spends keeping an eye out for a predator, is known to decrease when individuals of a prey species congregate in larger groups. This is due to the greater safety benefits of being in a group – greater risk detection (if one individual spots a predator then it is able to alert the rest of the group) and risk dilution (there is a lower probability of any one individual being targeted). The effect of group size on anti-predator vigilance is known to vary across species, and in some species group size has little or no effect.

Beauchamp developed a computer model that used a genetic algorithm to simulate predator-prey interactions while he changed several variables. One thing that the model demonstrated was that 'food patchiness' (how spread out food sources were) played a key role in how far group size impacted on anti-predator vigilance. Beauchamp found that, "Vigilance interacted with group size to [a] different extent depending on food patchiness level." (Beauchamp, 2017, p.309) The more spread out food sources were, the greater effect group size had on anti-predator vigilance. Of course, food isn't the only generalised concept playing a role in Beauchamp's theory. We can also see that he makes use of the concepts of predator and prey, both concepts that can be applied to an extremely diverse selection of animals.

The concept of food would have little value to a biochemist researching nutrition, but to the behavioural ecologist it is still a useful concept. While it is possible to provide a reductive explanation of food in terms of the micro and macronutrients of biochemistry, this level of detail is not necessary for the theories of the behavioural ecologist. Further to being unnecessary, an analysis of food at this level would be counterproductive. Breaking food down into its constituents would obscure nomological connections that are only apparent when we are dealing with more general abstract concepts like food.

Beauchamp aims to explain the behaviour of prey species in general. The model can be applied to all known species that are threatened by predators and could be applied to explain the behaviour of as yet undiscovered prey species. It could even be applied to alien ecosystems, if and when they are discovered. The diets of prey species are diverse, but in Beauchamp's model, they are unimportant and potentially distracting from the behaviour of prey species in general.

A similar point has been made by Fodor (1974 & 1997) when he argued for the autonomy of psychological laws. Although for present purposes it is not necessary to argue that the concepts of food or belief are irreducible, merely that in many instances it would be unhelpful or counterproductive to break down these generalised concepts to their more detailed lower-level parts[34].

While the concepts of food and, as I will argue in this chapter, belief still have scientific utility, as I have emphasised elsewhere in this thesis, integration does not mean that a concept will or should remain unchanged in the future. Like any scientific concept, an integrated BELIEF will remain open to revision as cognitive scientists refine their theories to fit existing and newly generated empirical data.

[34] As Jones (2004, p.412) argues - "no *particular* microphysical explanation can tell you that a certain macro-level generalization holds. But that doesn't tell you that there *can be no* microphysical explanation of the generalization"

It is possible that there will come a point at which the general concept of belief loses its utility within developmental psychology and fractures into two or more successor concepts. But as I will demonstrate in chapter **5**, such a fracturing remains a future possibility, rather than a present fact.

4.3 Infants mental understanding: rich or lean?

The scientific debate that I will focus on in the remainder of this chapter is whether infants' mental understanding abilities are rich or lean. In this section I will set the debate out in some detail before concluding the chapter with a commentary on the debate. I will show how rich theory has inspired a fruitful scientific research programme, while lean theory is unable to generate testable predictions.

In recent years numerous studies have shown that preverbal infants' behavioural expectations are sensitive to agents' false beliefs. There are two schools of thought as to the nature of the cognitive mechanism underlying this sensitivity. **Rich** theorists claim that the mechanism is metarepresentational, that is to say, infants form behavioural expectations by attributing a false belief (or proxy state) to the agent. **Lean** theorists claim that expectations are formed on the basis of behavioural rules, directly mapping an observed situation/behaviour with an expected behaviour.

In 1978 Premack and Woodruff published an influential paper entitled 'Does the chimpanzee have a theory of mind?'. The paper was not only influential because of its impact within the authors' discipline of primatology, but because it caused thinkers in other disciplines to seriously question what kind of cognitive processes underlie *human* mental understanding abilities.

In commentaries on Premack and Woodruff's paper, Daniel Dennett (1978) and Gilbert Harman (1978) suggested that a way of empirically testing whether a subject's understanding of the mind was metarepresentational or not was to test whether they had an understanding of **false belief**. Dennett suggested that a subject making a successful behavioural output prediction on the basis of a true belief would be indistinguishable from the subject making the prediction on the basis of the way the world was, rather than the agent's representation of it. So, for example, if both Hans and Celeste know that there is a sausage in the fridge and Celeste predicts that Hans will search in the fridge for the sausage, this prediction could either have been supported by her own belief that the sausage was in the fridge, or by an attribution of the belief to Hans that the sausage is in the fridge. But if a subject is able to make a correct prediction on the basis of a false belief, the subject must be representing the representations of the agent (which have the potential to misrepresent) rather than the actual state of the world. So if earlier in the day – unbeknownst to Hans – Celeste ate the sausage in the fridge, but still predicts that Hans will search for it there, we can infer that the prediction must have been made on the basis of an attribution of a belief, rather than on her own representation of the world.

A study was soon developed and carried out by Wimmer and Perner (1983) which became the blueprint for the traditional false belief task. In the task agent A places an object in a location X and leaves the scene, agent B enters the scene and transfers the object to location Y, thereby rendering agent A's belief about the location of the object false. The subject is then explicitly asked where agent A will search for the object when they return. To pass this task, child subjects are required to make a verbal prediction that agent A will search in location X.

In a meta-analysis of the traditional false belief literature taking into account 178 separate studies, Wellman, Cross and Watson (2001) concluded that children are unable to pass this test until around 44 months. The reason for the failure of children under 44 months was debated in the literature, with some maintaining that a conceptual shift to a full metarepresentational understanding of mind occurred at this age (e.g. Perner, 1991; Flavell *et al*, 1990) and others arguing that non-conceptual abilities prevented the effective deployment of metarepresentational concepts (Bloom & German, 2000) that were already present. All participants in the debate agreed, however, that in order to pass the traditional false belief task subjects must be able to effectively metarepresent the agent's belief.

The traditional false belief task relies on directly asking a question to the infant subject about where an agent will look for an object. But over two decades after the original traditional false belief task, Onishi and Baillargeon (2005) developed an entirely non-verbal false belief task, in which subjects were not explicitly asked where the agent would search. Instead the new non-verbal trial paradigm used looking times as a measure of violation of subjects' expectations of where an agent would search. The surprising results of this non-traditional false belief task showed that infants as young as 15 months passed the test by demonstrating an expectation that an agent with a false belief would search in the incorrect location. Following this initial finding, early competence has been demonstrated in numerous non-traditional false belief tasks (i.e. those tasks in which a child is not directly questioned) using a variety of different situations and experimental paradigms (e.g. Southgate *et al*, 2007; Song & Baillargeon, 2008; Kovacs *et al,* 2010).

This experimental data poses an interesting question, dubbed by Wilby (2012) as the **time lag problem**: why is it that infants show sensitivity to false beliefs in non-traditional false belief tasks as early as 7 months (Kovacs *et al*, 2010) and yet are unable to pass traditional false belief tasks until 44 months? (Wellman *et al*, 2001).

One potential answer to the time lag problem is to argue that the results of the non-traditional false belief tasks are best explained in terms of non-representational behavioural rules (Perner & Ruffman, 2005; Perner, 2010; Perner & Roessler, 2012). Infants using this strategy are able to pass non-traditional false belief tasks but are unable to pass the traditional false belief tasks until the point that they develop a full representational understanding of other minds, which comes with the ability to metarepresent, later in their development.

Following Caron (2009), I will refer to this explanation of early success on non-traditional false belief tasks as the **lean** theory. Opposing the lean theorists are the **rich** theorists, who argue that metarepresentation is necessary for success on both traditional and non-traditional false belief tasks. Below I summarise the two opposing positions.

Rich theory – Infants use metarepresentations much earlier than previously thought. Infants observe the state of the world (including, but not limited to, the behaviour of the agent) and attribute a representational state or states to the target agent on that basis. An expectation of future behaviour is inferred on the basis of this attribution[35].

Lean theory – Infants do not attribute representational states to the agent, but instead use a series of behavioural inference rules. Infants observe the state of the world and infer directly from this observation a prediction of the agent's behaviour.

Of course, if the rich theorists are correct, then they must appeal to other factors for an answer to the time lag problem.

4.31 A review of the empirical literature

In this section I will review some of the studies which demonstrate infants' sensitivity to false beliefs, briefly in the case of traditional false belief tasks and in a little more detail for non-traditional tasks.

Traditional false belief tasks

The original false belief task – devised by Wimmer & Perner (1983) – involves the transfer of an object to a different location, unseen by the agent, resulting in their possession of a false belief. In the study, children of varying ages watched a puppet show. In the show the puppet 'Maxi' placed a piece

[35] The nature and inferential richness of the state attributed to the agent by the infant is a matter of debate between rich theorists. I will not go into the specifics of the debate at this point. What is important presently is that a state is attributed and it is this that distinguishes the rich account from the lean account, which does not involve the attribution of any states.

of chocolate into the cupboard and then left the scene to play. While Maxi was away, his puppet mother moved the chocolate from the cupboard to a box. The subjects were explicitly asked "Where will Maxi look for the chocolate on his return?" The correct answer is obviously the cupboard. Maxi left the chocolate in the cupboard and was unaware that his mother had moved the chocolate in his absence. We should therefore expect Maxi to hold the false belief that the chocolate is in the cupboard where he left it, and that this false belief will drive his searching behaviour. The question was posed to three groups of children. Of the first group of children aged 3-4, none of the children gave the correct answer. In the second group aged 4-6, 57% gave the correct answer and in the final group aged 6-9, 86% gave the correct answer.

A variation of the traditional false belief task is the unexpected contents false belief task (Gopnik & Astington, 1988; Hogrefe, Wimmer & Perner, 1986; Perner *et al*, 1987). In this version of the task children are shown a Smarties[36] container and asked what they think it contains. The children respond that it contains Smarties. The tube is then opened to reveal that it actually contains pencils. The children are then asked what someone who has not seen the contents of the tube will say it contains. Generally, children under the age of 4 respond that they will say *pencils*. Interestingly, when asked what they initially thought was in the tube, children that fail the task will generally say that they thought it contained pencils.

A meta-analysis of traditional false belief tasks was performed by Wellman, Cross and Watson in 2001. It showed that success increased rapidly with age: at 30 months children have a 20% success rate, by 44 months this has increased to 50% and the percentage continues to increase from there.

Non-traditional false belief tasks

In this section I will first look at some of the experimental methods used in non-traditional false belief tasks. I will then describe three separate studies in order to give a sample of the variety of studies in the literature, firstly the original study by Onishi and Baillargeon (2005), secondly a study by Song and Baillargeon (2008), which demonstrates that infants are sensitive to beliefs caused by misleading appearances, and finally a study by Scott and Baillargeon (2009), which demonstrates that infants are sensitive to false beliefs caused by object identity.

While in the traditional false belief task it is just a matter of asking the subject what they think, "where will they look?" or "what will they think is in the tube?", non-traditional tests have to resort to less direct methods. Rather than being explicitly asked where the agent will look, non-verbal methods are used to discern the expectations of the subject. There are five experimental paradigms common in the literature and it is worthwhile to summarise these methods before describing some of the studies that employ them (adapted from Perner, 2012):

1) **Anticipatory looking –** The experimental set up generally involves the agent appearing in one of two places to look for or recover an object. Prior to the appearance of the agent, the Infant subject's eye gaze is tracked and taken as an indication of expectation of where the agent will appear or search. Examples include Clements and Perner (1994) and Southgate *et al* (2007)

2) **Violation of expectation –** Infants are shown two events, one that would be expected by someone with mature mental understanding abilities and another that would be unexpected. Looking for longer at one of the situations than the other is taken to mean that the infants' expectations have been violated. Examples include Gergely *et al*, 1995, Onishi & Baillargeon, 2005; Song *et al*, 2008. A more recent variation on the violation of expectation paradigm is the study by Moll *et al* (2017), who found that infants' expressed more tension in their facial

[36] A type of sweet that has milk chocolate contained in a crisp sugar shell

expressions when agents searched in the correct location for an object, which is taken to be a violation of expectation

3) **Helping –** The infant is tasked with helping the agent search for an object in one of two locations. If the agent is searching in the incorrect location on the basis of a false belief and the infant begins a search in the correct location this is interpreted as the infant subject understanding that the agent has a false belief. Examples include Buttleman *et al* (2009)

4) **Referential communication –** An object is hidden in a location and a transfer occurs out of sight of the agent. The agent then points at the location in which she believes that object to be and names it X. When subsequently tasked to recover X the infant recovers the item that the agent falsely believed was in the location pointed to. This is interpreted as the subject realising that the agent was referring to the object that she falsely believed was in the wrong location. Examples include Southgate *et al* (2010)

5) **Neural response –** The most recent experimental paradigm is to assess infants' neural response to an agent's behaviour. Using an EEG scanner, a study by Southgate and Veratti (2014) suggested that 14-month-old subjects predicted reaching behaviour where an agent had a false belief that an object was present, but did not make the same prediction where the agent had a false belief that an object was absent. A study by Kampis *et al* (2015) found that 8-month-old subjects represented an object, and then stopped representing that object when it disintegrated. However, they continued to represent the object where an agent had a false belief that it was still present.

Below I will describe three different types of non-traditional false belief tasks that involve some of the experimental paradigms described above. I will not question the validity of the five experimental paradigms described above, but it should be noted that they are not without their critics (e.g. Charles & Rivera 2009). However, as we shall see below, sensitivity to false beliefs is suggested by results from studies using all five of these diverse non-traditional methods. This lends credibility, as it reduces the probability that the results are false positives due to a particular experimental design. When taken together, the experimental results do seem very robust and it seems relatively safe to focus efforts on explaining the mental understanding abilities demonstrated by the infant subjects.

Unexpected transfer

The first non-traditional false belief task was conducted with 15-month-old infants in 2005 by Onishi and Baillargeon using the violation of expectation paradigm. The study involved four different trials, two of which involved situations designed to induce the infant to attribute true beliefs (TB) to the target agent, and two that were designed to induce the infant to attribute false beliefs (FB) to the target agent. The experiment involved the infants looking at the actions of an actor who stood in front of a scene containing a green box, a yellow box, and a toy watermelon slice. Between the actor and the scene there were two doors which, when open, enabled the actor to see the scene and to touch the objects within it. When the doors were closed, the actor was neither able to see the scene nor touch the objects.

All infants observed a pre-trial in which the actor was shown playing with the watermelon, placing it in the green box, and then subsequently retrieving the toy from the green box. During the main trial, infants once again saw the actor playing with the watermelon and placing it in the green box. Infants then saw one of the following four situations:

1) The actor watches as the yellow box moves halfway towards the green box and then returns to its original position. The actor has a true belief that the watermelon is in the green box

2) The actor watches as the watermelon slice makes its way from the green box to the yellow box. The actor has the true belief that the watermelon is in the yellow box

3) The doors are closed while, unobserved by the actor, the watermelon slice makes its way from the green box to the yellow box. The actor has the false belief that the watermelon is in the green box

4) The actor observes, through the open doors, the watermelon move from the green box to the yellow box. The doors then close and the watermelon returns to the green box. The actor has the false belief that the watermelon is in the yellow box.

After viewing one of the situations described above, half the infants were shown the actor reaching into the green box and the other half were shown the actor reaching into the yellow box.

The results showed that where the actor had a true belief (situations 1 & 2) that the watermelon was in a particular box, infants looked for longer when the actor searched in the other box. In the false belief situations (3 & 4) Infants looked longer when the actor searched in the box that the watermelon was actually in, i.e., not the box that an adult would expect the actor to believe the watermelon was in.

The results suggest that in the false belief condition, subjects had formed the expectation that the agent would search in the box which she or he falsely believed contained the watermelon.

Misleading appearances

In a 2008 study with 14.5-month-olds using the violation of expectations paradigm, Song and Baillargeon set out to discover whether infants would be sensitive to false beliefs caused by misleading appearances. In the familiarisation trial the actor agent showed a preference for reaching for a blue pigtailed doll rather than a skunk toy. The subjects and not the agent were shown two boxes, one of which had a tuft of blue hair attached to the inside of the lid, such that when the box was closed it would appear that one of the doll's pigtails was poking out of the box. In one of the conditions of the study, the subject observed as – out of sight of the agent – the doll was hidden in the non-tufted box and the skunk was hidden in the tufted box. When the agent arrived on scene subjects tended to look for longer where agents correctly searched in the non-tufted box, as opposed to when they searched incorrectly in the tufted box. The study suggests that infants had formed the expectation that the tufted box would mislead agents into falsely believing that it contained the preferred doll, rather than the undesirable skunk toy.

Object identity

In a 2009 study Scott and Baillargeon aimed to test if 18-month old infants were sensitive to false beliefs about object identity. The study involved two penguin toys, one of which divided into two pieces and one which didn't divide. By a series of familiarisation trials, infants were shown a scene in which the two-piece penguin was disassembled in one location and the indivisible penguin was in a separate location. An agent entered the scene, hid her key in the two-piece penguin, assembled it and left the scene. The object of the familiarisation trials was to enable the subject to recognise that the goal of the agent was to place the key in the two-piece penguin and that the two-piece penguin always started off in two pieces. In the false belief condition of the study, the agent entered the scene to find two covers, a transparent one and an opaque one. Underneath the transparent cover stood a fully assembled two-piece penguin. Subjects tended to look for longer when the agent reached for the penguin under the transparent cover than when they looked under the opaque cover. The agent had been shown to have the goal of reaching for the two-piece penguin. In the familiarisation trials the two-piece penguin had always started off in its divided state. It would be expected then that the agent, seeing that the penguin under the transparent cover was in one piece, would infer that the penguin under the opaque cover was the desired two-piece penguin. The results of the study suggest that this is indeed the expectation that the infants formed.

4.32 Interpretations of the data

In this subsection I will describe in detail how rich and lean accounts can successfully explain the data from the non-traditional false belief tasks described. I will do this as sympathetically as possible initially, but in subsection **4.33** I will show how the rich account has been at the centre of a rich research programme, while in contrast the lean account is unable to generate new empirical predictions. On this basis, the rich account should be favoured over the lean.

Rich interpretations of the data

A rich theorist explains success on non-traditional false belief tasks by supposing that infant subjects use a cognitive mechanism that enables them to attribute belief states to agents and predict future behaviour on the basis of this attribution. The metarepresentation forms an 'intervening variable' (Whiten, 1996) between the observed input and the expected output of a behavioural prediction.

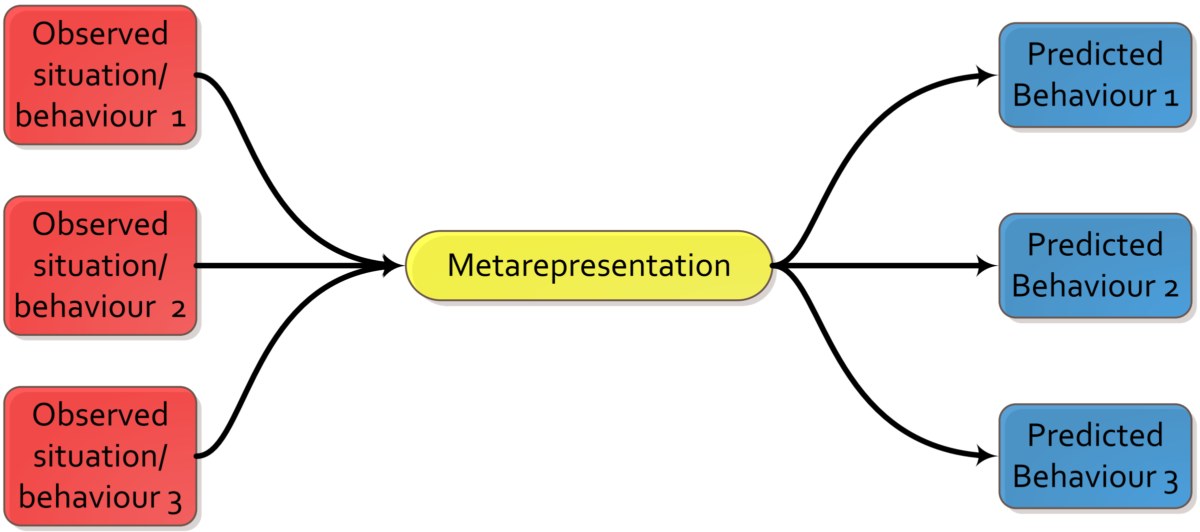
Observed situation/ behaviour 1
Observed situation/ behaviour 2
Observed situation/ behaviour 3
Metarepresentation
Predicted Behaviour 1
Predicted Behaviour 2
Predicted Behaviour 3

Figure 1. *A schematic of a rich account of behavioural prediction. A metarepresentation is attributed on the basis of an observed situation/behaviour, based on this attribution the subject now forms expectations of the agent's future behaviour*[37]

One of the best worked out descriptions of the belief attribution is that of Nichols and Stich (2003). In their account a model of the agent's beliefs is stored in a 'possible worlds box'. This is initially populated by the subject through a process of default belief attribution. This process involves the subject making a copy of all of their beliefs and attributing them to the agent. Essentially, the default position is that others believe what I believe and this will hold until evidence to the contrary becomes available. From the point of view of the subject, default belief attribution populates the possible world box with a set of true beliefs.

Deviation from default attribution is reliant on the subject's level-1 perspective taking mechanisms, which underlie sensitivity to the fact that from a particular position or orientation, an agent would be

[37] Note that this schematic depicts only a single metarepresentation; outputs would generally be the result of the interaction of several metarepresentations. For example, the expectation that the agent will search in a particular location for an object would involve not only a belief that the item was in a particular location, but also a desire to retrieve the object.

unable to see a particular object (Apperly, 2011)[38]. These mechanisms enable the subject to track the perceptual situation of the agent and to attribute perceptual states on this basis. Where a new perceptual situation, P, occurs which causes the subject to form a new belief, they are able to consult the output of the perspective taking mechanism and to attribute a perceptual state to the agent, either that of 'seeing that P' or 'not seeing that P' depending on whether the agent is judged to have received the belief causing perceptual information or not.

Once the perceptual state of seeing that P/not seeing that P has been attributed to the agent, the subject uses standing inference rules such as *seeing leads to knowing* and its opposite *not seeing leads to not knowing* to infer whether the new belief that they have formed is appropriate for the agent. If the agent is judged to have 'seen that P', then the new belief formed by the subject can be default attributed to the agent. If the agent is judged as 'not seeing that P' then default belief attribution is blocked and the original – now false – belief remains in the possible worlds box attributed to the agent.

Nichols and Stich (2003) admit that this process would not be so simple for situations in which a default attribution had not been made before it was apparent that the agent had a false belief. So, for example in the unexpected contents task, at the point where the subject believes that the tube contains Smarties, the agent was not present to default attribute this belief. In situations such as this, the subject must be able to attribute beliefs on the basis of the way the world appears to the agent. The agent sees the box of Smarties and on the basis of this would naturally believe that it contained Smarties. The subject does not believe that it contains Smarties, however, because they have seen the actual contents. Using the level-1 perspective taking mechanism, the subject is able to attribute 'not seen that P' where P is the opened container revealing the true contents, from which they infer that the agent has the normal belief about the contents of a Smartie box when confronted with one: that it contains Smarties.

Rich Interpretation of the Unexpected Transfer Task

Let's see how this model applies to the Onishi and Baillargeon (2005) non-traditional task. Take for example situation (iii). In familiarisation trials the agent plays with a watermelon in front of two boxes, one yellow, and one green. Once the agent finishes playing, they place the watermelon in the green box. Later they come to retrieve the watermelon from the green box and play with it some more, again returning it to the green box. All of these happenings are visible to both subject and agent and so, according to the rich account, the subject will hold in their possible worlds box a model of the agent's beliefs which match their own exactly. In the test trial the agent plays with the watermelon for a while before as usual returning it to the green box. However, a set of doors close, blocking the agent's view of the boxes. While the agent's view is blocked, the subject observes as the watermelon moves from the green box to the yellow box. On the basis of this observation the subject will then update their own beliefs, removing the belief that the watermelon is in the green box and replacing it with the belief that it is in the yellow box. The subject is also sensitive to the fact that the perceptual situation that enabled them to form a new belief would not have been visible to the agent. A perceptual state of 'not seeing' is attributed to the agent. From this attribution and the standing inference rule *not seeing leads to not knowing*, the subject's newly formed belief 'the watermelon is in the yellow box' is blocked from being default attributed to the agent, and the belief 'the watermelon is in the green box' remains part of the subject's model of the agent's beliefs in the possible worlds box.

[38] Level-1 perspective taking abilities are to be contrasted with Level-2 perspective taking abilities in which subjects are sensitive not just to whether the agent did or did not see a particular object, but to the fact that they may see the same object in a different way (Apperley, 2011).

When the agent reappears from behind the doors, the subject forms the expectation that they will search for the watermelon in the green box on the basis that they have the false belief that that is where the watermelon currently is. When the agent goes to the yellow box to successfully retrieve the watermelon this expectation is violated and infant subjects look longer at this situation than they do when the agent searches for the watermelon in the green box, in accordance with their false belief.

Rich interpretation of misleading appearances

The rich account of attribution is different when applied to the misleading appearances task. Unlike the unexpected transfer task, there is not a point in time at which the subject and the agent share a belief, i.e., one that has been default attributed by the subject to the agent due to a shared belief forming experience. In this task, observing that the doll is placed in the non-tufted box, the subject forms the belief that it is in the non-tufted box.

When the agent arrives on the scene, level-1 perspective taking mechanisms enable the subject to infer that the agent was not exposed to the belief forming perceptual situation. This leads to the subject attributing a 'not seen that P' perceptual state to the agent. Following the *not seeing leads to not knowing* rule, the subject infers that the agent has not formed the true belief that that the doll is in the non-tufted box. As the agent did not see the doll placed, they must base their beliefs about the location of the doll on the information that is available to them. The most obvious indicator of the whereabouts of the doll is the protruding tuft and so the subject attributes the belief that the doll is in the tufted box, as this is the way that the situation would appear to the agent.

Perhaps it could be objected that the attribution of a false belief is not necessary to explain the data, and that all that is required is that the subject doesn't attribute a true belief about the doll's whereabouts. In other words, the agent is searching in ignorance. However, if this were the case we would expect the agent to adopt a random search strategy and this expectation would not be violated if the agent searched in either box. The experimental results predicted by this no attribution account would be that the infant agent would look equally long at situations in which the agent searched in the tufted and the non-tufted box, and this is not the case.

Rich interpretation of object identity

In the rich interpretation of this task the subject must attribute to the agent the false belief that the visible penguin is not the desired two-piece penguin, but is in fact the undesirable indivisible penguin. Having seen the two-piece penguin assembled and placed under the transparent cover, while the indivisible penguin is placed under the opaque cover, the subject forms the true belief that the visible penguin is the desired two-piece penguin. When the agent enters the scene, the subject is able attribute 'not seen that P', where P is the placing of the penguins under their covers, and from this infers that the agent does not know which penguin is under which cover. In the absence of this knowledge, the only information available to the agent as to the identity of the visible penguin is its undivided state. During the familiarisation trials, on entering a scene the two-piece penguin was always in its divided state whereas the indivisible penguin was always obviously undivided. On the basis of the information available, the false belief that the penguin under the transparent cover is the indivisible penguin and the corresponding false belief that the desired two-piece penguin is under the opaque cover is attributed to the agent. On the basis of these attributions, the subject forms the belief that the agent will search under the opaque cover.

Lean interpretations of the data

It has been argued by Josef Perner and others (Perner & Ruffman, 2005; Perner, 2010, Perner & Roessler, 2012) that infants show sensitivity to false beliefs, not because they attribute metarepresentations to agents, but because they possess a set of behavioural rules, which link observed behaviour with expected behaviour directly, rather than involving any 'intervening variables' (Whiten, 1996).

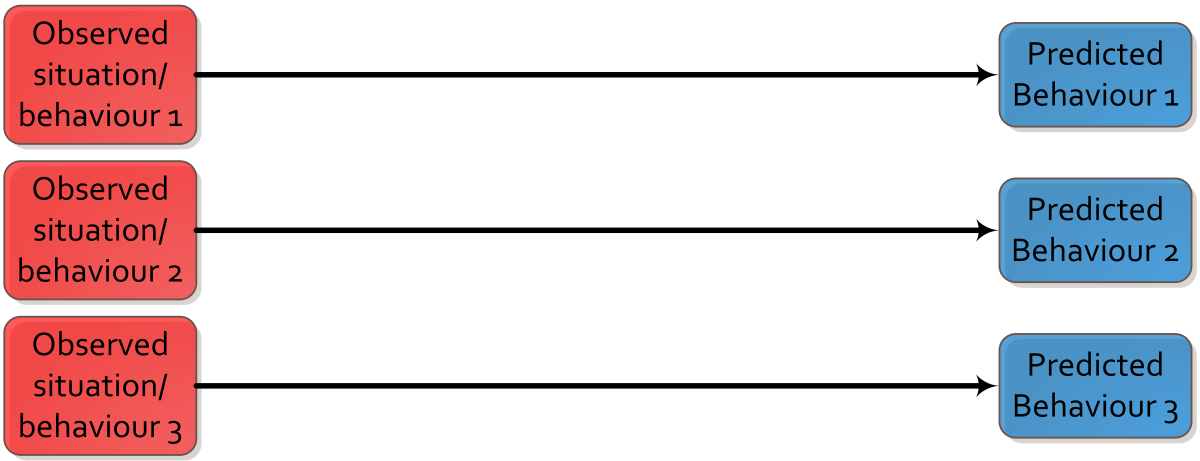
Observed situation/ behaviour 1
Predicted Behaviour 1
Observed situation/ behaviour 2
Predicted Behaviour 2
Observed situation/ behaviour 3
Predicted Behaviour 3

Figure 2. *A schematic of a lean account of behavioural prediction. Expected behaviour is inferred directly from observed situation/behaviour.*

The argument is similar to one from a related parallel debate in the comparative psychology literature. Working with non-human primates, Tomasello and others performed a series of experiments (Hare *et al*, 2000; 2001; Tomasello *et al* 2003; Melis, 2006) which the authors interpreted as demonstrating that the chimpanzee subjects of the studies were able to metarepresent the state of 'seeing'. A group led by Daniel Povinelli argued that data is equally well explained by behavioural rules. Take the example of an experiment in which chimpanzees are shown to preferentially signal for food to non-blindfolded humans over blindfolded humans (Povinelli & Vonk, 2004), the rich account would have it that from the fact that Mary has her eyes covered while Suzy has them uncovered, the chimp attributes 'seeing' to Suzy and not to Mary. The chimp chooses to gesture for food to Suzy as only humans that can 'see' respond to gestures. Povinelli and Vonk's alternative account reaches the same destination, but misses out what they see as the unnecessary middle step. The lean interpretation of this result is that the chimp observes that Mary is blindfolded whereas Suzy is not. The chimp chooses to gesture to Suzy because only humans with their eyes uncovered respond to gestures. Povinelli and Vonk argue that this strategy of simply removing the middle step of metarepresentation attribution can be applied to a rich interpretation of the results of any of the studies which purport to show that apes are able to metarepresent 'seeing', without any loss of explanatory power or predictive accuracy.

Perner takes this strategy and transplants it into the debate over the nature of the mechanism underlying infants' success on non-traditional false belief tasks. He formulates a "recipe for posing Povinelli's challenge" (Perner, 2010), which involves letting a rich theorist specify inference rules of the form:

Observation > metarepresentation > behaviour

And then collapsing it into a behavioural rule of the form

Observation > behaviour

Above I described how a rich interpretation could be applied to three different types of non-traditional false belief task. In this section I will apply Perner's recipe to these interpretations to see if the lean theory is able to explain the same range of data.

Unexpected transfer

In this task the subject forms the expectation that the agent will search in the green box for the object. In the rich interpretation the agent was default attributed the belief that the object was in the green box because both subject and agent were able to observe as the agent placed the object into the green box. In the rich interpretation described above, the chain of inference looked something like this:

Agent observes object move to the green box

> Agent believes that object is in green box

> Agent will search for object in green box.

Applying Perner's recipe to this chain of inference we get:

Agent observes object move to the green box

> Agent will search for object in green box.

The resulting general behavioural rule being that: 'agents look for objects in the last location they saw them'. Indeed, this was the behavioural rule suggested by Perner and Ruffman (2005) in response to Onishi and Baillargeon's (2005) original non-traditional false belief study. The lean theorist claims that the expectation that is violated, as indicated by the longer looking time, is formed on the basis of this behavioural rule.

Misleading appearances

The 'agents look for objects in the last location they saw them' behavioural rule used by the lean theorist to explain the results of the unexpected transfer task does not work with this task as the agent does not see the object placed in a location. For this task the rich theorist has the subject's inference chain looking something like this:

Agent observes box that appears to contain sought object (tufted box)

> Agent believes that object is in the tufted box

> Agent will search in tufted box for object.

Applying Perner's recipe to this chain we arrive at:

Agent observes box that appears to contain sought object (tufted box)

> Agent will search in tufted box for object.

This results in the general behavioural rule: 'agents will search for an object in locations that appear to contain that object'.

Object identity

In this task the subject forms the expectation that the agent will search under the opaque cover for the divisible penguin. On the rich interpretation the subject's inference chain took the following form:

Agent observes object that does not appear to be the sought object

> Agent believes that object is not the sought object

> Agent will search in different location for sought object.

Once again, removing the middle step we arrive at the following inferential steps:

Agent observes object that does not appear to be sought object

> Agent will search in different location for sought object.

And the general behavioural rule that we extract is: ‘agents will search in an alternative location for a desired object when presented with an object that does not appear to be the desired object’.

4.33 Rich or lean?

Perner’s (2010) method of generating a lean interpretation by removing the metarepresentational step from the rich interpretation brings us seemingly to an impasse. No experimental design is able to conclusively demonstrate that a subject is using a metarepresentation rather than a behavioural rule. To decide between the two camps then, we need to look to other factors.

While it may be conceded that Perner’s recipe is effective in generating behavioural rules to explain the full variety of results coming out of the non-traditional false belief literature, it has been argued that the lean account makes for a poor theory, as the post-hoc nature of Perner’s method means it is unable to produce empirical predictions.

By working on the theory that infants are attributing representations to agents to achieve their mental understanding, rich theorists have been able to make novel predictions and design experiments to generate data that confirms their predictions. The nature of the rich account allows rich theorists to make specific, testable predictions, which so far seem to be consistent with the wealth of research being conducted in the field. In their 2017 review, Scott and Baillargeon state that, “To date, over 30 published reports using non-traditional tasks have provided positive evidence of false belief understanding in infants (under age 2 years) and toddlers (age 2–3 years).” (Scott & Baillargeon, 2017, p.238).

Onishi and Baillargeon’s 2005 original non-traditional study tested infants’ understanding of false beliefs about the location of an object. Since this experiment, understanding of false belief has been demonstrated where those false beliefs are about a variety of different states of affairs. For example, Scott and Baillargeon (2009) and Buttellmann *et al* (2015) demonstrate infants have an understanding of false beliefs about an object’s identity; Scott and Baillargeon (2010) and Moll *et al* (2017) show they have an understanding of false beliefs about the properties of objects; while Choi and Luo (2015) show infants have an understanding of false beliefs about puppet’s personalities.

More recent experiments have also shown that false beliefs can be attributed on the basis of a variety of different types of information (or disinformation). In the Onishi and Baillargeon study the infants attributed false beliefs when an agent was aware of a state of affairs, but then that state of affairs changed when the agent left the scene. Other experiments have since shown that infants can attribute false beliefs on the basis of misleading appearances, for example, Scott and Baillargeon (2010) and Moll *et al* (2017), while Buttelmann *et al* (2014) showed false beliefs can be attributed on the basis of reality not following a previously observed pattern.

As we can see from the above, the rich theory has inspired an extremely active and fruitful research programme. The progress that has been made in the area of non-traditional false belief tasks since the original Onishi and Baillargeon (2005) paper has been impressive, and the gamut of data now available would have been unlikely to be generated if it were not for the rich account.

As for the rival lean account, as Carruthers points out it is “entirely parasitic on positive results provided by the infant-mindreading hypothesis [what I am calling the rich account]” (2013, p.150). The very nature of Perner’s method means that it needs to be applied to a pre-existing explanation consisting of a chain of inferences involving the attribution of a metarepresentation. Indeed, that is exactly what I did in 4.32, above, in order to generate behavioural rules that enabled the lean account to handle the range of data from the three types of experiment I had described.

Nor has any lean theorist, whether discussing ape or infant, attempted to refine their theory by suggesting a range or set of behavioural rules subjects are likely to possess together, due to common features or a likely order of acquisition. The theory is thus left with enough wriggle room to enable it to fit any pattern of empirical data. For any study showing infant mental understanding abilities, the lean theorist can use Perner's method to generate a behavioural rule. Where a study shows a deficit in mental understanding compared to a mature subject, the lean theorist can simply claim that the subject does not yet possess the relevant rule.

As a result of this under-specification and its post-hoc nature, the lean theorist is unable to produce any testable predictions of their own (Fletcher & Carruthers, 2013), leaving their theory unfalsifiable by future empirical research.

The fruitful research programme inspired by the rich interpretation gives it an advantage over the lean. There are several other considerations which arguably favour the rich account. In terms of simplicity the lean account may initially seem to have an advantage. The account does the same work without needing to posit metarepresentations. However, as the empirical work on the subject has continued to expand, the lean theorist needs to posit more and more behavioural rules to explain the data.

While these considerations do not rule the lean theory out, it would make it appear to be somewhat stubborn to stick to the reactive and post-hoc lean account. Indeed, the majority of theorists working in this area over the past few years accept some kind of metarepresentational account, while adherents of the behavioural rules account are few and far between. This is typical of a deteriorating research programme.

For present purposes it is not necessary to decide between the rich and lean theories, but it is clear that the rich theory has the edge over the lean. At this stage, the important point to take away from this case study is that many theorists hold that infants are attributing metarepresentations to agents in order to form expectations about their future behaviour. As mentioned earlier in the chapter, theorists such as Southgate and Vernetti (2014) use the term 'belief', but are reluctant to make commitments about the nature of the belief that is being attributed and its relation to belief "traditionally characterised". Yet others such as Apperly and Butterfill (2009) argue that what is being attributed is not a belief, but rather a state that acts as proxy for a belief. Despite these reservations, attempting to prove or disprove the theory that infants attribute some kind of belief-like state has inspired the development of a series of increasingly ingenious experiments.

However, even if there were agreement on the rich side that the representational state being attributed by infants was a belief, would this show that the concept of belief still played a useful role in the cognitive sciences? As discussed in section 4.1 I argue that it wouldn't. The belief that is being attributed by the subjects of the experiments described above plays an indirect, rather than a direct role in the developmental psychologists' theories. In this indirect sense, the theoretical status of belief in developmental psychology is much like the concept of the hunting spirit from the hypothetical anthropological study that I imagined earlier in this chapter. It is something that experimental subjects use, but that does not mean that we must take it seriously as a scientific concept. It would not follow that the concept of belief still had scientific utility[39]. Indeed, it is uncontroversial that at least in some circumstances adult humans attribute full beliefs, but this should not be used as a defence of BELIEF as a useful scientific concept.

[39] One could potentially object to my relegation of belief in this sense to an indirect role in developmental theories by arguing that by taking belief seriously, subjects make it real. Such an argument could appeal to idea of "mindshaping" (Zawidzki, 2013) in which our belief-desire interpretative framework is not only a tool for understanding minds, but also plays a significant role in regulating the ways that our minds function. This is not a line that I will develop here.

Rather than focus on the indirect role of belief in the theories of developmental psychologists, I will argue that a generalised concept of BELIEF is playing a different, and direct, role. In the next chapter I will show that the technical term "attribute" should be identified with the generalised concept of belief that I described in chapter 2.

5. Beliefs about beliefs

I have two aims in the final chapter. Firstly, in section **5.1** I will show that a generalised concept of belief plays a primary role in the theories of developmental psychologists in debate over whether infant mental understanding abilities are rich or lean. This will demonstrate that the concept of belief has been integrated into at least one area of cognitive science and therefore still has scientific utility.

Secondly, in section **5.2** I will examine a potential point at which the BELIEF concept could fracture – the distinction between beliefs and subdoxastic states. I will question the viability of this distinction by reviewing evidence which suggests that inferential integration is a spectrum rather than a binary property. Finally, I will show that the most recent work in developmental psychology describes a functional role of infant beliefs about beliefs that is hard to categorise into sub-belief states.

5.1 BELIEF as a direct concept in developmental psychology

I established in the previous chapter that, while not conclusive, the rich theory should be favoured over the lean. You will recall that earlier I characterised the rich theory as follows:

Infants use metarepresentations much earlier than previously thought. Infants observe the state of the world (including, but not limited to, the behaviour of the agent) and attribute an intentional state or states to the target agent on that basis. An expectation of future behaviour is inferred on the basis of this attribution.

The core claim of this section is that terms like 'attribute' and 'ascribe' are technical terms for the generalised concept of belief that I described in chapter **2**.

To say that:

The subject attributes a belief that X to an agent

is equivalent to saying that:

The subject has the belief/believes that the agent has a belief that X

Demonstrating this identification provides the evidence required to show that the concept of belief, albeit without the 'belief' label, has been smoothly integrated into at least one area of cognitive science and that it plays a key direct role in contemporary theories of human mental understanding abilities. This is enough to show that the concept still has scientific utility and is not currently a suitable candidate for elimination, as claimed by Jensen (2016).

To make this identification, I will analyse the role that 'attribute' and similar technical terms play in developmental psychologists' theories. As an illustrative example consider the basic outline of the unexpected transfer task that I discussed in the previous chapter (from Onishi & Baillargeon, 2005):

Unexpected transfer task – A subject watches as an agent plays with a toy watermelon slice and places it inside a green box. The agent leaves the scene. Out of sight of the agent, but observed by the subject, the watermelon moves from the green box to a yellow box. The agent then returns to the scene in search of the watermelon.

The results of the experiment show that, above chance, 15-month-old subjects form the expectation that the agent will look in the green box for the toy watermelon slice, rather than the yellow box which actually contains the toy watermelon.

It is the task of developmental psychologists to explain the results of this experiment and others like it. While most rich theorists agree on the basics of the mechanism that explains the results, the descriptions and terminology that they use are various and often less than clear.

In the discussion section of their original paper, Onishi and Baillargeon (2005 p257) say that, "These results suggest that 15-month-old infants already possess (at least in a rudimentary and implicit form) a representational theory of mind: They realize that others act on the basis of their beliefs and that these beliefs are representations that may or may not mirror reality."

Victoria Southgate (2013, p.5) says that "…preverbal infants appear to represent another person's representation of the world" and that infants are "representing the world from the perspective of the other" (2013, p.10).

Carruthers (2013a) argues that infants possess a mechanism that "provides infants with the concepts and core knowledge necessary to represent the mental states of other agents, of all basic types (including beliefs that are false and appearances that are misleading)." (p155). In the same paper he also uses the term **ascribe –** "infants will have only a limited number of cues that they can use for ascribing thoughts to other agents, at least at the outset" (p160) and uses this interchangeably with the term **attribute –** "What is present at the outset is just a capacity to attribute propositional representations of various basic types, including belief." (p143).

'Attribute' is the term used by Scott and Baillargeon in their recent review paper on 'Early false belief understanding'. They frame their question as "When are children first able to attribute false beliefs and other counterfactual mental states to others?" (Scott & Baillargeon, 2017, p.237).

As we can see, developmental psychologists use a variety of technical terminology to explain the results of false belief tasks. But despite the different ways of describing the process, the **minimal rich claim** on which all of these theorists would agree is that, based on perceptual input, an infant subject generates a representation of the target agent's representation of the watermelon in the green box. The subject then forms their expectation that the agent will search in that location on the basis of this metarepresentation.

I stated at the start of this section that saying that an infant 'attributes' a belief to an agent is equivalent to saying they have a belief about the agent's belief. In that statement I focused on the technical term 'attribute', but I take it that 'attribute' and the other technical jargon used by theorists (ascribe, metarepresent, etc.) are different ways of making this common minimal rich claim.

Clearly, the infant subject's metarepresentation in the minimal rich claim is playing the functional role of the generalised concept of belief that I set out at the end of chapter 2. The metarepresentation is a state that has a state of affairs as its content (in this instance that state of affairs is the representational system of the agent), it has been caused by the subject's perceptual input, and on the basis of this state, the infant subject forms an expectation of the agent's behaviour.

If we were to attempt to translate the developmental psychologist's explanation of the experimental evidence into everyday terms, the most natural translation would be: the infant expects that the agent will look in the green box because they **believe [or think]** that is where the agent **believes** the watermelon to be.

Why does the developmental psychologist prefer to employ technical terminology rather than everyday terms? The developmental psychologist might think that the ordinary language explanation that employs the term 'belief' says more than they want it to. As we saw in chapter **2** terms like 'think' or 'believe' have very specific uses in English (dependent on context and conversational setting) and use of this term has a whole range of potential connotations that they want to avoid.

As I noted at the end of chapter **3**, even the four elements of the functional role of belief that Schwitzgebel (2011) gives in the Stanford Encyclopaedia of Philosophy, which I reproduced in chapter **1**, have several unwanted connotations. If the developmental psychologist were to say that the infant 'believes that the agent believes that X', someone might infer from 4) that the infant should be able to verbalise that belief, or at least verbalise the expectation that was formed on the basis of that belief. But of course, we are dealing with preverbal infants so such a conclusion would be false. We know from the results of the traditional false belief task that infants are unable to verbalise their expectation of where the agent will look until three or four years of age. A further complication may be caused by the use of the term 'reflection' in 1), as this may be taken to imply that reasoning involving these metarepresentations is necessarily a reflective conscious process. As we discussed in chapter 1, the extensive use of the term 'typically' in Schwitzgebel's list suggests that none of the four points are necessary conditions, but nonetheless the list is too evocative of the special emphases and assumptions regularly found in folk usages of 'belief' and belief-like terms.

To maximise the clarity of their theory, developmental psychologists want to avoid any unnecessary commitments, real or perceived, about infants' metarepresentations. They don't want to make any claims about the nature of this state or its broader functional role further to that described in the minimal rich claim above. When making this minimal rich claim it is better to avoid any potential connotations derived from usage of an everyday language term and to use a newly co-opted technical term such as 'attribute'.[40]

The concept of belief, albeit under a different label, is therefore playing a key role in developmental psychologists' theories explaining the behavioural expectations of subjects in non-traditional false belief tasks.

It should also be clear that the role that the concept of belief is playing in the minimal rich claim is direct, rather than indirect. The claim is that the infant is attributing a representational state to the agent. The representational state attributed by the infant/subject plays an indirect role in the theory as described in the previous chapter, but the theory builder is making the claim that the infant/subject attributes and so 'attribute' (or any of the equivalent technical terms described above) is playing a direct role in the theory. I have argued that 'attribute' and other technical terms are equivalent to 'believe' and if this identification is correct, the concept of belief is playing a direct role in the minimal rich claim.

The virtue of the concept of belief in the context of the rich versus lean debate is its generality, the theorist need not make any overly specific claims about the nature of the state beyond that directly warranted by the experimental results. The category of belief may contain a variety of sub-states with more specific functional roles, but in this case the general concept allows the theorist to explain the experimental results without having to commit prematurely to any particular one.

Employing the general concept of belief means that the minimal rich claim is compatible with a number of different theories about the specific functional role of infants' metarepresentations and can be used as an agreed framework within which scientists can attempt to establish which of these is correct.

The rich theorists then are in agreement that infants are metarepresenting, rather than employing behavioural rules. This rules out one potential answer to the time lag problem that I mentioned earlier, but the rich theory does not itself provide an answer.

If the minimal rich claim is correct and preverbal infants of 15 months are able to pass non-traditional false belief tasks by metarepresenting an agent's representational states (or as I am claiming having a belief about an agent's representational states), then why is it that verbal infants younger than 44

[40] Or 'ascribe' or 'metarepresent'.

months are unable to pass traditional false belief tasks? Why is it that the metarepresentational state that enables infants to form expectations about an agent's behaviour is not able to inform their verbal responses until much later?

This is a question I will return to in the next section. For now, what is important is that the debate is taking place within the framework of a commonly accepted minimal rich claim in which the generalised concept of belief plays a direct role. While the concept of belief plays a direct role in the minimal rich claim, I have also explained why developmental psychologists have abandoned the label 'belief' in favour of technical jargon without the same folk connotations.

5.2 Doxastic and subdoxastic: division or continuum?

One potential way that belief could fracture is into **beliefs proper** and **subdoxastic states**. Many of the examples that Jenson (2016) uses to show the potential for fracture in the concept of belief are experiments which use implicit tests to measure subjects' attitudes. The Implicit Association Test (IAT), for example, shows that individuals that explicitly declare that they have egalitarian beliefs can nonetheless show an implicit bias against certain sections of society.

Jenson's contention is that explicitly asking an individual their beliefs about a certain section of society and their taking part in the IAT are two methods of detection which detect two distinct states. This demonstrates that the category of belief has fractured, that it contains more than one type of state, and that these different types of state are best treated by science as distinct. An umbrella term that is used to refer to both of these states becomes redundant and is pragmatically eliminated from cognitive science.

In this section I will argue that the distinction between beliefs and subdoxastic states is far from clear cut. And that substantial overlaps between the two categories exist.

It is not my intention here to conclusively demonstrate that no such distinction can be drawn. Rather, it is my intention to demonstrate that on the current state of scientific knowledge, we are unable to draw a clear distinction with any level of confidence and that it is far from clear that it is inevitable that such a distinction will be drawn in the future.

In subsection 5.21 I will look at several ways in which the category of belief could fracture. In 5.22 I will describe evidence that suggests that inferential integration is a matter of degree, rather than a binary property. In 5.23 I will look at the most recent theories on the nature of infants' metarepresentations and how these describe a functional role that can't be captured by the belief/subdoxastic state binary distinction.

5.21 Proposed fractures in BELIEF

In an important 1978 paper, Stephen Stich argues that there is a folk distinction between beliefs proper and what he calls subdoxastic states. He thinks that cognitive science has ignored this distinction, and that it would be better served to incorporate it.

Beliefs, he notes, can be inferred from other beliefs, but to prevent an infinite regress or circularity, it cannot be the case that all beliefs are inferred from other beliefs. So some beliefs must be non-inferential. "The class of non-inferential beliefs marks the boundary between two different sorts of psychological states. On one side of the boundary are beliefs, both inferred and non-inferential. On the other side are the psychological states which, though not beliefs, are part of the casual process leading to belief formation." (Stich, 1978, p.501).

Stich gives as an example the rules of grammar. Any competent English speaker is able to distinguish grammatical from non-grammatical sentences, yet unless they had received formal training in grammar, they are not able to verbalise how they judged the sentence as grammatical or not.

We have a belief about whether a sentence is grammatical or not, but that belief was not inferred from other beliefs. It is a non-inferential belief that has been generated by a body of information that itself does not qualify as a belief. The rules of grammar, at least for non-grammarians, are subdoxastic.

Stich proposes that we can categorise beliefs and subdoxastic states by considering whether the state is accessible to consciousness and how inferentially integrated the state is. "…[B]eliefs form a consciously accessible, inferentially integrated cognitive subsystem. Subdoxastic states occur in a variety of separated special purpose cognitive subsystems." (Stich, 1978, p.507-508).

In the example of the rules of grammar, the subject is not able to consciously access their subdoxastic states. While these subdoxastic states form the basis for judgements about whether a sentence is grammatical or not, they are not integrated with our other beliefs: "they are inferentially impoverished, with a comparatively limited range of potential inferential patterns via which they can give rise to beliefs and a comparatively limited range of potential inferential patterns via which beliefs can give rise to them." (Stich, 1978, p.507)

We can apply this distinction to an example of the IAT that I discussed earlier. Recall that participants, even those that had avowed egalitarian beliefs, often nonetheless revealed that they more easily associated negative words with dark-skinned faces. It seems that while the participants in IAT studies had egalitarian beliefs, they may still have subdoxastic states that are prejudiced against certain groups. The belief/subdoxastic distinction is therefore seemingly useful in explaining the results of the IAT. Two incompatible states can exist within a single agent because while the state that drives the explicit verbal behaviour of the agent is a belief, a state that is inferentially integrated, the state that drives the implicit behaviour revealed by the IAT is a subdoxastic state, which is insulated from the rest of cognition.

Stich's distinction between beliefs and subdoxastic states broadly maps to a more modern distinction between type-1 and type-2 cognitive processes, a distinction that has been posited in several fields across cognitive science (e.g., see the papers in Evans & Frankish, 2009). The precise details about the nature of these processes are not universally agreed upon and vary between different accounts. There is also disagreement as to whether the type-1/type-2 processes map to distinct cognitive systems that correspond to each process type. The properties of system 1 and system 2 have been summarised in a helpful table by Frankish and Evans (2009, p.15), which I reproduce below:

System 1	System 2
Evolutionarily old	Evolutionarily recent
Unconscious, preconscious	Conscious
Shared with animals	Uniquely (distinctively) human
Implicit knowledge	Explicit knowledge
Automatic	Controlled
Fast	Slow
Parallel	Sequential
High capacity	Low capacity
Intuitive	Reflective
Contextualised	Abstract
Pragmatic	Logical
Associative	Rule-based
Independent of general intelligence	Linked to general intelligence

In addition to Stich's belief/subdoxastic distinction, several other theorists have proposed dual process style distinctions between types of belief or between beliefs and other sub-belief or belief like states. Daniel Dennett for example suggests a distinction between belief and opinion (Dennett, 1991). For Dennett, beliefs are states which enable us to form expectations about the behaviour of others, as the preverbal infant subjects do in the non-traditional false belief task. Opinions, however, are limited to language users and are "…essentially bets on the truth of sentences in a language that we understand" (Dennett, 1991, p.143). Dennett further goes on to make an eliminativist prediction, similar to the type of elimination suggested by Jenson (2017) many years later:

> "My empirical hunch is that a proper cognitive psychology is going to have to make a sharp distinction between beliefs and opinions, that the psychology of opinions is really going to be rather different from the psychology of beliefs, and that the sorts of architecture that will do well by, say, nonlinguistic perceptual beliefs (you might say "animal" beliefs) is going to have to be supplemented rather substantially in order to handle opinions."

Other suggestions as to how belief should splinter have been proposed by Frankish (2004), Gendler (2008a; 2008b), and Sterelny (2003).

5.22 A spectrum of integration?

As we have seen in the previous section there are several ways in which distinctions can be drawn within the range of the category of belief. I don't have the space to provide specific arguments against each of the individual accounts that I mentioned in the previous sections. Rather, in this section I will focus on one important way in which states within the category of belief have been distinguished – inferential integration.

I am not claiming that differences along this dimension are essential to each and every possible distinction between states within the category of belief, but it is certainly an important one. As we saw in the previous chapter, it is key to Stich's distinction between beliefs and subdoxastic states. It also features heavily in Gendler's distinction between beliefs and aliefs. For Gendler, aliefs are dispositional, associative, and to some extent innate[41], "our dispositional aliefs depend on the associational patterns that have been laid down in our minds as the result of our experiences and those of our genetic ancestors." (Gendler, 2008, p.651).

My aim in this section is not to prove beyond doubt that there can be no robust distinctions between states within the category of BELIEF in principle, but rather to make us pause for thought. If inferential integration is a spectrum, then we should not be so hasty in pronouncing the elimination of BELIEF. To explain empirical results, scientists may need to posit states that do not fit comfortably within any of BELIEF's putative successor categories. In these instances, BELIEF has an important role to play in scientific theory construction. This role could well be temporary, as a placeholder while the precise role of the state is investigated and correctly categorised. But it could also be that the belief-like states that cognitive scientists posit do not fit into neat categories and that BELIEF is a category that is here to stay.

One indicator of whether a state is inferentially integrated is its **propositional content responsiveness**. A characteristic of inferentially integrated beliefs is that they can be combined with other beliefs to infer new beliefs.

For example, Hans is trying to think of a suitable present for Celeste's birthday. He knows that Celeste enjoys growing fruits in her small garden. Hans is told by a mutual friend that Celeste does not have any citrus plants in her garden. Hans takes a cycle to the local garden centre and sees a lemon tree

[41] Ignoring the complications around this term that we explored in chapter 1

for sale. From his belief that Celeste doesn't have any citrus plants and his belief that lemon is a type of citrus plant, Hans is able to infer that Celeste doesn't already have a lemon plant in her garden and decides to purchase one for her birthday.

Already existing inferentially integrated beliefs can also be updated or removed in light of new propositional content. Continuing with our example, Hans has seen an advertisement on social media that the garden centre is offering free delivery on all annual plants. Hans, believing that the lemon tree was an annual plant, had inferred that the lemon tree was eligible for free delivery. He took the lemon tree to the counter, paid, and then asked for the free delivery service. The garden centre worker promptly informed him that the lemon tree was actually a perennial plant, not an annual plant. Hans immediately grasps the inferential implications of this new information. His belief that the lemon tree is eligible for free delivery is incorrect and he begrudgingly pays the small delivery charge.

In contrast to beliefs, subdoxastic states are not inferentially integrated in this way, and we would not expect them to be responsive to propositional content in the same way that beliefs are. Indeed, their lack of content responsiveness is precisely why they seem to give a neat explanation of how seemingly contradictory states can coexist within a single agent.

However, recent papers from Eric Mandelbaum (2016), Sophie Stammers (2017) and Ferguson et al (2019) discuss a range of evidence which suggest that what we would consider subdoxastic states can be sensitive to propositional content.

As Stammers (2017) points out, showing that what she calls implicit attitudes (equivalent to what we have been calling subdoxastic states) are sensitive to propositional content is on its own not enough to show that the we can't draw a distinction between categories of states on this basis. It could be that there are two distinct clusters of states, with one cluster (e.g., subdoxastic states) showing a relative insensitivity to propositional content and the other (e.g., beliefs) showing a higher sensitivity to propositional content. If there is no overlap between the two clusters, then it would still be possible to argue that they form two distinct kinds of state.

In order to rule out this possibility, we can demonstrate overlap by describing examples of cases in which beliefs proper fail to update in the face of new and contradictory propositionally structured content.

Implicit attitudes are sensitive to propositional content

A study by Briñol et al (2009) and discussed in Mandelbaum (2016) showed that the strength of an argument can have an impact on implicit attitudes as measured by the IAT. Briñol et al presented two groups with different written arguments in favour of hiring more black academics. The first group were given a set of strong arguments, including the idea that hiring more black academics would increase the quality and number of academics without having to increase tuition fees and that the number of students in each class could be reduced, thereby increasing the quality of education. The second group were given a set of weak arguments in favour of hiring more black academics, including the idea that hiring more black academics would mean that the university would be part of the national trend towards hiring more black academics and that doing so would allow the current academics more free time.

The results of the experiment showed that the group that received the stronger arguments demonstrated more of a positive response to dark-skinned faces in an IAT taken after reading the arguments. The difference was more marked in a 'high elaboration' condition in which subjects were induced to more carefully consider the arguments.

This experiment suggests that whatever type of state is driving the implicit behaviour measured by the IAT, they are at least to some extent sensitive to a single presentation of propositional arguments in written form.

Other experiments point to the same conclusion. A study by Van Dessel *et al* (2015) asked participants to imagine two groups of peoples, the Niffites and the Luupites. Participants were also given details of the types of name typical to members of each of these two peoples. Luupite names always included two consecutive vowels (e.g. Loomalup & Ageelup) while Niffite names always included two consecutive consonants (e.g. Borrinif & Kennunif). Half of the participants were then told that in the next stage of the experiment, they should approach individuals with typical Luupite names and avoid individuals with typical Niffite names. The other half of the participants were told the reverse.

Participants were then given an IAT test where they had to pair positive or negative words with typical Luupite and typical Niffite names. After the IAT, participants were then asked to explicitly rate their warmth towards individuals with typical Luupite and Niffite names on a scale of 1 to 9. The results of the experiment showed that the participants that had been asked to approach individuals with Niffite names and avoid individuals with Luupite names, showed a preference for Niffite names on both the implicit and explicit tasks. The opposite was true for the group in the other experimental condition, where they were told to avoid individuals with Niffite names and approach individuals with Luupite names.

Another study by van Dessel *et al* (2019) found that implicitly measured attitudes towards Mahatma Ghandi were changed by a single reading of a negative text[42]. The study included three separate experiments using three different ways of measuring implicit attitudes. In two of the three experiments, participants that had read the negative Ghandi text revealed more negative attitudes towards Ghandi, as measured by the implicit tests.

As with the Briñol et al (2009) experiment, the two Van Dessel et al (2015; 2019) experiments show that a single presentation of a propositionally structured stimulus was enough to change implicitly measured attitudes. This provides direct evidence against the idea that the state that drives those implicit attitudes is an inferentially insulated subdoxastic state.

You could argue that, rather than the state that drives implicit attitudes being directly sensitive to propositional content, the changes observed in the IAT are driven by changes to explicit attitudes, which then in turn impact on implicit attitudes. For example, you could argue that in the Briñol et al (2009) experiment, the subjects' explicit attitudes are changed as a result of the argument in favour of black academics, and this caused the participant to have a series of positive thoughts paired with thoughts of black academics. This pairing of positive thoughts with black academics could then have provided a form of internal association training, and it is this, rather than the propositional content, that then caused the changes observed in the IAT.

A further study by Van Dessel et al (2016) aimed to rule out the possibility that the effect on implicit attitudes was mediated by changes in explicit attitudes. They repeated their 2015 Niffite/Luupite methodology, but this time gave participants a list of character traits for the two groups of peoples. Participants were told that "…one group 'are very good people; they are peaceful, civilized, benevolent, and law abiding, whereas the other group 'are very bad people; they are violent, savage, malicious, and lawless."

[42] The negative text included details of how "due to religious reasons, Gandhi refused the use of modern medicines on his sick wife who died afterward, yet allowed the same treatment on himself and recovered when he fell ill shortly thereafter" the neutral text described Ghandi living in South Africa for a time (Van Dessel et al, 2019, p.268). The negative text presents a historical viewpoint that is contested.

Participants that had been told that the Niffites were bad and the Luupites were good, but had been told to approach the Niffite individuals and avoid the Luupites, tended to rate the Niffites badly on the explicit rating, but still showed a slight preference for them on the IAT. The opposite was true in the reverse condition. This dissociation showed that changes in implicit attitudes were affected directly by the instructions to approach or avoid, and were not being affected by participants' explicit attitudes.

In a review article on the evidence that implicit attitudes can be rapidly updated in the face of new, propositionally structured information Ferguson *et al* (2019, p.333, citations in original) argue that:

> "…there is currently not enough evidence to support a two-system or two-process approach that assumes different learning constraints for implicit versus explicit impression updating (Amodio & Ratner, 2011; De Houwer, 2014; see also Ferguson et al., 2014; Ruisch, Cone, Shen, & Ferguson, 2018). Instead, the evidence is more thoroughly consistent with perspectives positing that propositional thinking can strongly impact implicit cognition."

They do, however, suggest that a lot more empirical work needs to be done on the topic to establish the nature of the mechanisms underlying implicit attitudes, and do not rule out the possibility that evidence for a dual process account will be uncovered in the future.

Beliefs proper can fail to integrate

Having established that implicit attitudes are sensitive to propositionally structured content, I now move on to describe examples of cases in which beliefs proper can fail to update despite a subject being presented with new and directly contradictory evidence.

Stammers (2017, p.1842) quotes David Lewis (1982, p.436) as saying:

> "I used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel… So each sentence in an inconsistent triple was true according to my beliefs, but not everything was true according to my beliefs."

We can all probably think of an anecdote from our lives in which we held contradictory beliefs, or when we failed to update some of our existing beliefs in light of new evidence that directly contradicted them. As a young lad my friend's parents left him home alone while they went out for the evening. He alerted me to this fact and I called round for a visit. We left the house on a mission to acquire alcohol and cigarettes. On our return, the drive which had been empty when we left was now filled with his parent's distinctive van, the very same van that they had left in. Not fully appreciating the implication of this, and flush with the success of our mission, we leant against the van to smoke a cigarette and plan what we would do with our parent free evening.

Examples of this phenomena can be found in the psychological literature. We can see it in the well-established psychological phenomenon of **belief perseverance**. Several early studies used a debriefing paradigm in which fabricated evidence was presented to participants, in order to induce inaccurate beliefs. Participants were subsequently debriefed and informed that the evidence that supported their new beliefs was fabricated. Tests showed that the new beliefs persisted in participants, despite the debriefing.

For example, in a study by Ross *et al* (1975) participants were given 25 cards on which were printed one genuine and one fictitious suicide note. The participants were asked to identify which of the two was genuine. One group of participants was told that they had performed above average on the task, one group was told that they had performed at average, and another group was told that they had performed below average. After a delay, participants were then given a full debriefing. It was explained to them that their performance (above average/average/below average) had been assigned to them before the task had taken place, and their reported performance was in no way indicative of their ability to distinguish genuine from fake suicide notes. Despite the debriefing, in a subsequent questionnaire individuals that had been previously told that they were above average tended to rate their abilities to distinguish between genuine and fake suicide notes more highly than those individuals that had been told that they had performed at average or below average.

In a similar study Anderson *et al* (1980) presented participants with fabricated data that suggested that individual firefighters' job performance correlated with their approach to risk taking. One group of participants was given data suggesting a positive correlation, while the other group was given data which suggested a negative correlation. As in the previous experiment, participants were fully debriefed and were told that the evidence of a correlation was fabricated. In subsequent tests, the participants showed that their belief in the fabricated correlation persisted, for example by estimating that successful firefighters were more (or less) likely to take risks.

In addition to belief perseverance, the phenomenon of preference reversal seems to suggest that we can simultaneously hold contradictory beliefs. Bortolotti (2009) discusses a classic study by Tversky *et al* (1988) in which a group of participants were told that there are around 600 road traffic accident fatalities each year in Israel. In experiment 1, a group of participants were then presented with the following data on two different programs to control road traffic accident fatalities (adapted from Tversky *et al,* p.373):

	Expected number of casualties	Cost
Programme A	570	$12m
Programme B	500	$55m

Participants were then asked to advise the Minister of Transportation which of the two programs is preferable. 67% of participants in this group stated that program B was preferable, despite its higher cost. In experiment 2, a second group of participants were then presented with the same information except that the cost of program B was not included. The participants from this group were then asked to suggest a value for the cost of program B, which would mean that the two programs were equally preferable. 96% of participants chose a value for program B lower than $55m, suggesting that they would prefer program A if given the full details.

Bortolotti (2009, p.80) interprets these findings as demonstrating that individual participants have contradictory beliefs about which programme the Ministry for Transportation should adopt:

> We believe that the Minister should implement programme B in order to save the lives of 70 more people a year, even if the programme costs $43 million more that programme A. We also believe that the Minister should implement programme A, which would save fewer lives, unless programme B costed considerably less than $55 million.

In this section I have looked at evidence which suggests that states underlying the implicit attitudes measured in the IAT and other implicit measures can be sensitive to propositional content and therefore can be to some extent inferentially integrated. I have also reviewed evidence which suggests that beliefs proper can fail to update in light of new contradictory evidence and can also coexist with contradictory beliefs in a single cognitive system. This evidence suggests that beliefs proper can be inferentially insulated.

Recall that in the evidence reviewed by Jenson (2016) that I discussed in chapter **1**, one of the purported advantages of fracturing belief into distinct states was that it allowed us to explain how the results of the IAT show that some individuals seem to simultaneously hold non-racist and racist beliefs. That explanation was that while the non-racist egalitarian views were beliefs, which are inferentially integrated, the racist beliefs were actually subdoxastic and inferentially insulated. However, if there is the potential for overlap between states on the spectrum of inferential integration, then any explanatory advantage of positing two states evaporates.

Again, to emphasise, this section is not intended to show that it is impossible in principle to identify distinct states within the generalised concept of belief. What it is intended to show is that on the current state of scientific knowledge, we are not able to identify distinct functional roles with any confidence. That is not to say that we should rule out our scientific knowledge advancing to the point that we are able to identify these distinct roles. Equally, at this stage we can't rule out the possibility that we will be unable to identify distinct functional roles within the generalised concept of belief. It may be that functional roles are fluid, and that any one individual state can play a variety of different functional roles depending on the cognitive context that obtains at any particular time.

On the current state of our scientific knowledge, we are unable to straightforwardly assign states to categories within the generalised category of belief. And so it is premature to abandon this category.

5.23 Problem states – infants' metarepresentations

In this subsection I will discuss one type of state that doesn't fit easily within the distinctions that have been proposed to fracture the generalised concept of belief – infants' metarepresentations. While on the current state of evidence, the precise functional role has yet to be established, the balance of evidence seems to support an account of a single system developing over time into adulthood, as opposed to a dual-system solution. The functional role of infants' metarepresentations in the single system account is hard to characterise into the sub-belief categories that we explored in subsection 5.21. Further work needs to be done to establish the precise functional role of these states. While it may be that further empirical and theoretical work will enable cognitive scientists to confidently categorise infants' metarepresentations as one or another type of sub-belief categories, it is also a possibility that further work could reveal a range of states that is too rich and fluid to be captured by a simple binary classification.

I mentioned earlier that rejecting the lean theory and accepting the minimal rich claim does not provide a satisfactory answer to the time lag problem[43]. To answer that question, we need to explore further the functional role of infants' metarepresentational states and how they interact with the rest of cognition.

Within rich theory, two types of account have been proposed that offer a solution to the time lag problem. The first of these is a two-system account that would broadly fit with the project of fracturing belief into subcategories. The second is a single system account, in which infants' metarepresentations become increasingly integrated and accessible over time as the infant develops improved processing capabilities. I argue in this subsection that the evidence supports this latter account, and that, if correct, the functional role of metarepresentations in this account make them difficult to categorise into the post-belief successor categories that we explored in section 5.1.

[43] Why is it that infants show sensitivity to false beliefs in non-traditional false belief tasks as early as 7 months and yet are unable to pass traditional false belief tasks until 44 months?

Below, I briefly set out the two accounts and how the evidence weighs in favour of the single-system account. I then move on to discuss the implications of this for BELIEF.

One potential answer to the time lag problem is a two-system solution, for example, that proposed by Apperly & Butterfill (2009). According to this account, early success on non-traditional false belief tasks is achieved using an early developing system that is efficient but inflexible. Children are not able to pass the traditional false belief tasks until around 3 or 4 years of age, which is when a separate, flexible, yet effortful, belief reasoning system develops. The two systems continue to exist and operate in parallel into adulthood. Apperly and Butterfill (2009, p.957) summarise their account as follows:

> "…[L]ater-developing theory-of-mind abilities involve flexible cognitive processes that, by their very nature, depend on language or executive function, whereas infants' precocious abilities are underwritten by a distinct set of cognitively efficient processes that do not depend on language and executive function."

According to this account, the metarepresentations underlying early success on non-traditional false belief tasks have limited inferential integration and are not consciously accessible to the subject, while those of the later system are fully inferentially integrated and consciously accessible. If such an account were correct, this would strengthen the case for fracturing beliefs into subcategories, as the two types of metarepresentation involved in the two systems would seem to divide along something broadly like Stich's belief/subdoxastic state distinction.

If this were the case, while BELIEF has proved useful in the rich vs lean debate, that usefulness may be relatively short-lived. The eliminativist could argue that at best it is a placeholder concept, which would be useful just so long as it took cognitive scientists to accurately categorise posited BELIEF states into the appropriate successor categories.

However, in the past few years, the weight of evidence seems to support a picture of a developing functional role that does not map neatly to any sub-distinctions within the category of belief. These accounts (e.g., Carruthers, 2016; Setoe *et al,* 2017) argue that humans have a single system underlying false belief understanding and that this system steadily develops from infanthood onwards.

These processing limitation accounts argue that the resources that underpin preverbal success on non-traditional false belief tasks are sufficient to pass traditional false belief tasks, but that the failure on these tasks observed until around 3 to 4 years is due to the added processing demands of traditional false belief tasks, compared to the non-traditional tasks. Only once language and executive function have developed sufficiently are infants able to successfully deploy their conceptual resources and pass traditional false belief tasks.

In their recent review, Scott and Baillargeon (2017) review the evidence in favour of the processing limitations account over the two-system account.

Firstly, counting against the two-system account is evidence from neuroimaging. Apperly and Butterfill (2009) appeal to number cognition as an area in which it is well established that there are two separate systems, one efficient and inflexible and one effortful and flexible, both of which persist into adulthood. If a two-system solution evolved for number cognition, it makes it seem more likely that a similar two-system solution developed for social cognition. However, neural imaging evidence counts against this analogy: studies in adults show that the different number cognition systems involve distinct brain areas (e.g. Hyde & Spelke, 2011), while later studies showed that there is substantial overlap in the brain regions involved in traditional and non-traditional false belief tasks (e.g. Bardi *et al,* 2017).

Secondly, Apperly and Butterfill's account predicts that the inflexible system will show what they call "arbitrary signature limits" (2009, p.957) to infants' metarepresentational reasoning. However, the sheer variety of non-traditional false belief tasks that have been devised – and which infants have successfully passed – have demonstrated that infants' false belief reasoning is far more flexible than predicted by the two-system account.

Thirdly, supporting the processing limitations account is the fact that the age at which infants are able to pass the traditional false belief task has been substantially reduced using new techniques to reduce the processing demands of the task. This is a result that is predicted by the processing limitations account but not by the two-systems account. For example, a recent study by Setoe *et al* (2017) featured several additional steps to the traditional location-switch false belief tasks. The additional steps were designed to give "…children practice at interpreting a 'where' question and producing a response by pointing at one of two pictures." (Scott & Baillargeon, 2017, p.245). If these practice trials were included, 2.5 year old infants performed above chance on this traditional style task. A clear demonstration that failure on early traditional tasks was due to processing demands, and not the limitations of an inflexible early developing system.

The evidence, then, favours a single system account over the two-system alternative. Accepting the single system account makes it difficult to categorise the metarepresentations that underlie success on non-traditional false belief tasks. infants' beliefs about beliefs are clearly not language involving as they can be demonstrated in 15-month-old preverbal infants. They also cannot be said to be accessible, as children fail when asked to verbalise their beliefs in traditional false belief tasks. These considerations would suggest that infant metarepresentations might be best categorised as subdoxastic on Stich's (1978) belief/subdoxastic distinction.

Yet the range of behaviour measured in non-traditional false belief tasks is not limited to what are arguably implicit measures, such as the violation of expectation or gaze direction experiments. For example, the study by Buttleman *et al* (2009), using the helping paradigm, shows that 18-month-old infants are able to perform complex intentional actions like opening a locked box on the basis of the output of their false belief reasoning.

Recent evidence (Király *et al,* 2018) suggests that 3-year-old infants can form beliefs about beliefs retrospectively, based on episodic memories and not just current perceptual input. A finding which Carruthers (2018, p.11352) argues shows that: "Far from being inflexible and automatic, it seems that young children's mindreading abilities can make use of executively controlled searches of episodic memory and/or working-memory-demanding inferences drawn on the basis of such episodic memories."

Also, if the single system account is correct, infants' metarepresentational systems (in combination with improved processing abilities) allow older children to pass traditional false belief tasks, and also form the basis of the adult system in which it would be uncontroversial to say that subjects have fully inferentially integrated, accessible and linguistic beliefs about agents' beliefs.

Infants' beliefs, while showing some properties of what Stich would call subdoxastic states, also show properties of what he would call beliefs. Infants' beliefs seem to cut across proposed sub-belief categories throughout development.

Even in adulthood, beliefs about beliefs may continue to defy binary classification. Work by Schneider *et al* (2012) show that adults automatically form beliefs about agents' beliefs even when performing a cognitively demanding task unrelated to belief tracking. These automatically generated beliefs were measured by implicit eye tracking measures, but during a debrief participants were shown not have been explicitly reasoning about agents' beliefs[44]

[44] Five participants suspected of explicit belief reasoning were excluded from the study.

Carruthers (2016) reviews the evidence from both infant and adult studies, and develops an account of a single belief reasoning system that originates in infancy and gradually develops over time to fully mature in adulthood. The mature system is able to operate in different modes: “This system can be used in ways that do, or do not, draw on executive resources (including targeted searches of long-term memory) and/or working memory (such as visually rotating an image to figure out what someone else sees).” (Carruthers, 2016, p.141).

If this single system/two modes account is correct, then it suggests that in adulthood the functional role of beliefs about beliefs can be fluid, that in one mode they are capable of being fully inferentially integrated and accessible, yet in another they are inferentially insulated and inaccessible to consciousness.

In this section I have explored some of the sub-belief categories into which the generalised concept of belief has been said to fracture. However, in subsection 5.22, we reviewed evidence which showed that inferential integration, a line along which BELIEF has been claimed to fracture into sub-belief states, may be more of a spectrum rather than a clean distinction. In subsection 5.23, I explored the most recent empirical and theoretical work on the precise functional role of infants’ metarepresentations. I found that these metarepresentations show characteristics that straddle proposed post-BELIEF distinctions. Furthermore, evidence from adults suggests that the functional role of metarepresentations may be fluid, sometimes functioning in an inferentially integrated and consciously accessible mode, and under other circumstances operating in an inferentially insulated and consciously inaccessible mode.

In this section I give just a rough outline of the experimental work and theoretical debates that aim to clarify the precise functional role of infants’ metarepresentations and their relationship to the metarepresentations of older children and adults. However, it is clear that there is far more work to be done before a scientific consensus on this functional role emerges. It may be that the consensus that emerges confirms or refines proposed post-belief successor states. In this case, a generalised concept of belief will continue to play an important role in cognitive science until the point such a consensus occurs. However, it may be that future work in cognitive science reveals a fluid spectrum of functional roles that cannot be categorised into simple binaries. In this case, BELIEF may continue to play an important role further into the future, as an umbrella term for this range of states.

Conclusion

In this thesis, I have assessed the current and future role of the concept of belief in cognitive science.

In **chapter 1** I assessed an eliminativist challenge to the concept of belief. I rejected that challenge because I found that the robustness/fragility framework that Jenson (2016) used to assess whether belief should be eliminated was unconvincing. While Jenson's method was unconvincing, I did accept that there is a reason why the concept of belief faces a threat of elimination. Clearly as science progresses, any theory must be open to refinement. The fracturing of belief is one potential refinement that scientists might deem it necessary to make, and this raises the possibility of conceptual redundancy and elimination.

I set out a new framework in which to judge the eliminativist challenge to belief based on the utility test – is it useful to distinguish things of kind K from things that are not of kind K? If there is no use in distinguishing a category, then that category should be eliminated. One of the key lessons learned from the case study of innateness was that elimination can be discipline specific. There may be a use for a concept in some disciplines and not others. As cognitive science is an interdisciplinary endeavour, then it is possible that the category of belief may have utility in some areas but not others. It could be the case, then, that in some disciplines within cognitive science, the concept of belief would be pragmatically eliminated while it continues to serve a purpose in others.

To begin to evaluate whether the concept of belief should be eliminated or not, in **chapter 2** I moved on to examine the concept itself. 'Belief' is a term that has a wide and varied usage in the English language. It is reasonable to question what possible scientific value such a multi-faceted concept could have. However, in this chapter I showed that the concept of belief as it features in belief-desire psychology is not simply to be identified with the usage of the term 'belief' in the English language. Rather, the concept of belief is doubly theoretical. Theoretical in the sense that it is part of the core theory that underlies our mental understanding abilities, and also theoretical in the sense that it has been reconstructed by abstracting away from everyday usage of 'belief' and belief-like terms.

I showed that uncovering the philosophical concept of belief from everyday usage faces three major obstacles. Firstly, the term 'belief' has a variety of specialist uses in everyday discourse. Secondly, the concept of belief is often left unverbalised. Thirdly, BELIEF is often packaged together into a more complex concept and referred to by a belief-like term, giving it connotations over and above the basic concept itself. I demonstrated how philosophers like Davidson (2006 [1963]) abstracted away from this complexity to reconstruct the core of belief-desire psychology, identifying BELIEF as one of the two key concepts that we use to rationalise human actions. I also showed how Davidson would have been comfortable with the label of 'belief' for his core concept because of the history of its usage in epistemology. I finished the second chapter by setting out the functional role of belief.

In **chapter 3** I considered two rival accounts of what a *belief* is – representationalism and dispositionalism. I suggested that by embracing dispositionalism, at least in the recent form in which it is developed by Schwitzgebel (2002; 2003; 2013), we might have a way of avoiding the eliminativist threat. I set out the representationalist and the dispositionalist accounts of *belief* before evaluating the two positions. I showed how the representationalist account was able to supply the basis for detailed and flexible causal explanations of output, while the dispositionalist account could not. Moreover, the dispositionalist account could still not avoid the traditional objections of the argument from interactive function and the argument from the absence of behaviour. The two arguments offered in favour of the dispositionalist account, that it better fits with our intuitions and that it is better able to account for cases of 'in-between' believing, were shown to not favour one account over the other. I concluded by rejecting dispositionalism. While representationalism is vulnerable to the eliminativist threat, it provides a far better account of *belief*.

In **chapter 4** I began to lay the groundwork for applying the utility test to the concept of belief. I describe the way in which concepts can play direct and indirect roles in scientific theories, arguing that only by showing that belief plays a direct role in scientific theories would we be able to conclude that belief had scientific utility. I then demonstrated that even though a general concept may be reducible to lower-level concepts, it does not mean that it does not have scientific utility in certain disciplines.

The remainder of chapter 4 is taken up with describing a case study of a recent debate in developmental psychology about the nature of the mechanisms underlying preverbal infants' mental understanding abilities. Rich theorists argue that preverbal infants' mental understanding abilities are achieved by a system of metarepresentational states that model a target agent's cognitive system, while lean theorists argue that infant subjects form expectations on the basis of a set of behavioural rules. I go into detail on how each of the rival accounts explain the experimental data, concluding that the rich account is to be preferred because it has inspired a rich research programme in which different aspects of infants' mental understanding abilities have been explored in an increasing variety of ingenious experiments. In contrast, the lean theory makes no predictions, leaving it unfalsifiable and its interpretations of data are entirely ad hoc and reliant upon pre-existing rich interpretations.

In **chapter 5** I argued that the concept of belief plays a direct role in rich theories of infant cognition, but that the concept had been relabelled with new technical terminology to prevent confusion with folk usage. This is a clear demonstration that the generalised concept of belief passes the utility test in at least one discipline of cognitive science. A concept that can be shown to have scientific utility should not be eliminated. I noted that the virtue of the concept of belief in this debate was its generality. The minimal rich claim was that infant subjects were attributing a belief to agents, something I argued was equivalent to saying that infant subjects have beliefs about the agent's belief. As the precise cognitive role of infants' beliefs about beliefs has yet to be established, then to avoid making unwarranted claims, developmental psychologists need to make use of a general concept when constructing their theory.

One view could be that the generalised concept of belief is just a temporary placeholder: a general concept that can be used in the early stages of theory construction to be replaced by a more specific concept when one has been developed. If this were the case, once belief has been fractured into successor states, the concept of belief could be swapped out for a successor state with a more specific functional role that matched better the functional role of infants' metarepresentations. On this view eliminativists may be premature in their declarations of the elimination of belief, but they would ultimately be proven correct.

An analogy could be made with the concept of a planet. The planet concept started off as an observational concept used to refer to points of light observed in the night sky that behaved differently to stars. Through the development of telescopes and, more recently, space probes we have learned more and more about planets – their orbits, their composition, and their atmospheres. Our scientific concept of planet was gradually sharpened up as more and more empirical data was generated. Eventually, we saw the category of planet fractured. In 2006 the International Astronomical Union settled upon a commonly agreed definition that relegated Pluto, something that we had previously considered a planet, to the category of 'dwarf planet' and as the prototype of a new class of Trans-Neptunian Objects known as Plutoids (IAU, 2006).

However, that a concept has been fractured does not mean that the general concept automatically becomes redundant. The example of a general concept such as food, as discussed in chapter 4, shows that even where a general concept has been fractured, it can still play a useful role in theories at higher levels of abstraction. This may be the case with the general concept of belief.

In chapter 5 I explored some evidence which suggests that the fracturing of belief may not be as straightforward as some eliminativists assume. I looked at one proposed way in which belief could fracture, into beliefs proper and subdoxastic states. I reviewed evidence that inferential integration, the line along which the proposed fracturing occurs, may be more of a spectrum than a binary property. This evidence does not show that it impossible to fracture beliefs in this way, just that it has not yet been achieved and the result is not a foregone conclusion.

As an example of a kind of state that crosses proposed fracture lines, I highlighted the example of infants' metarepresentations. The evidence suggests that these states are not language involving and are in some respects insulated from the rest of cognition, but also that they guide complex intentional behaviour and that they can be attributed on the basis of executively controlled searches of memory. Even into adulthood, evidence suggests that beliefs about beliefs occupy a variety of different functional roles depending on the cognitive context. While this is preliminary evidence and far from conclusive, if it were the case that the functional roles of beliefs can be fluid, then there would be a continuing need for a generalised concept of belief.

We saw from the case study of innateness that we can save a label without saving the concept. For the concept of belief though, I would argue that we have saved the concept but not the label. The concept as it features in developmental psychology is not the belief of everyday English usage, but the doubly theoretical concept abstracted by philosophers over the past few decades. We needn't be too attached to the label 'belief'. Another thing that we learned from the case-study of innateness is that using the same label both for a folk concept and a scientific concept can cause confusion, especially when communicating with non-specialists. It seems sensible then that new technical terms have come to replace the potentially misleading term 'belief'.

The concept of belief, though not labelled as a 'belief', currently plays a useful, direct role in cognitive science. The concept is a useful tool and one should not discard a useful tool before it has been replaced. It isn't clear when or even if the concept will be fractured into sub-states. Even if a fracturing occurred, the concept of belief could continue to have scientific value as a high-level concept.

When and if the general concept of belief becomes redundant should be determined by scientists following the empirical data. Only when the concept becomes redundant should we consider it eliminated. If we accept this, modest integrationism should be the default position of scientifically minded philosophers. While the concept of belief may not be fashionable, it is certainly still serviceable.

Bibliography

Anderson, C. A., Lepper, M. R. & Ross, L. (1980) ‘Perseverance of social theories: The role of explanation in the persistence of discredited information’ *Journal of personality and social psychology*, vol. 39, no. 6, p. 1037-1049

Apperly, I. A. (2011) ‘Mindreaders: The cognitive basis of Theory of mind’ *Psychology Press*, Hove

Apperly, I. A, & Butterfill, S. A. (2009) ‘Do Humans Have Two Systems to Track Beliefs and Belief-Like States?’, *Psychological Review*, vol. 116, no. 4, p. 953-970

Ariew, A. (1996) ‘Innateness and Canalization’, *Philosophy of Science*, vol. 63 p. S19-S27

Astuti, R. & Harris, P. L. (2008) ‘Understanding mortality and the life of the ancestors in rural Madagascar’ *Cognitive Science*, 32 (4). p. 713-740

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R. & Brass, M. (2017) ‘Brain activation for spontaneous and explicit false belief tasks overlaps: new fMRI evidence on belief processing and violation of expectation’ *Social Cognitive and Affective Neuroscience*, vol. 12(3) p. 391-400

Barone, P., Corradi, G. & Gomila, A. (2019) ‘Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis’ *Infant Behaviour and Development*, vol.57

Baumeister, R. & Bushman, B. (2008) ‘Social Psychology: Human Nature’, Belmont: Thomas-Wadsworth

Beauchamp, G. (2017) ‘The spatial distribution of foragers and food patches can influence antipredator vigilance’, *Behavioural Ecology*, vol. 28, **1**, p. 304-311

Bloom, P & German, T. (2000) ‘Two reasons to abandon the false belief task as a test of theory of mind’ *Cognition,* ***77***, p. 25-B31

Bortolotti, L. (2009) ‘Delusions and other irrational beliefs’ Oxford University Press

Botterill, G. & Carruthers. P. (1999) ‘The philosophy of psychology’ Cambridge University Press

Boyd, R. (1991) ‘Anti-Foundationalism and the Enthusiasm for Natural Kinds’, *Philosophical Studies*, **61**, p.127-148

Briñol, P., Petty, R. E., & McCaslin, M. J. (2008) ‘Changing Attitudes on Implicit versus Explicit Measures: What Is the Difference?’ In R. E. Petty, R. H. Fazio, and P. Briñol (Eds.) ‘Attitudes: Insights from the New Implicit Measures’ New York: Psychology Press

Buttelmann, D., Carpenter, M. & Tomasello, M. (2009) ‘Eighteen-month-old infants show false belief understanding in an active helping paradigm’ *Cognition*, **112**, p. 337-342

Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015) ‘What you get is what you believe: Eighteenmonth-olds demonstrate belief understanding in an unexpected-identity task’ *Journal of Experimental Child Psychology*, 131 p. 94-103

Butterfill, S. A. & Apperley, I. A. (2013) ‘How to Construct a Minimal Theory of Mind’ *Mind & Language*, Vol. 28, **No. 5**, p. 606-637

Caron, A. J (2009) ‘Comprehension of the representational mind in infancy’ *Developmental Review*, **29**, p. 69-95

Carruthers, P. (1996) 'Simulation and Self Knowledge: A Defence of Theory-Theory' in Carruthers, P & Smith, P.K.(eds.) (1996) 'Theories of Theories of Mind' Cambridge University Press. p 22-38

Carruthers, P. (2006) 'The Architecture of the Mind' Oxford University Press

Carruthers, P. (2011) 'The Opacity of Mind: An Integrative Theory of Self-Knowledge' Oxford University Press

Carruthers, P. (2013a) 'Mindreading in Infancy' *Mind & Language*, vol 28, **2**, p. 141-172

Carruthers, P. (2013b) 'On Knowing Your Own Beliefs: A Representationalist Account' in Nottleman, N. (ed.) (2013) 'New Essays on Belief' Palgrave MacMillan p. 145-165

Carruthers, P. (2016) 'Two systems for mindreading?' *Review of Philosophy and Psychology*, vol 7, p. 141-162

Carruthers, P. (2018) 'Young children flexibly attribute mental states to others' *Proceedings of the National Academy of Sciences of the United States of America* vol 115 (45), p. 11351-11353

Chalmers, A. (2011) 'Drawing Philosophical Lessons from Perrin's Experiments on Brownian Motion: A Response to van Fraassen' *The British Journal for the Philosophy of Science*, vol. 62, **4**, p. 711-732

Charles, E. P. & Rivera, S. M. (2009) 'Object permanence and method of disappearance: looking measures further contradict reaching measures' *Developmental Science* 12:6 p. 991-100

Choi, Y. & Luo, Y. (2015) '13-Month-Olds' Understanding of Social Interactions' *Psychological Science*, vol. 26 (3) p. 274-283

Churchland, P. M. (1981) 'Eliminative materialism and the propositional attitudes' *Journal of Philosophy*, **78**, p. 67-90

Clements, W. A. & Perner, J. (1994) 'Implicit Understanding of belief' *Cognitive Development*, **9**, p. 377-395

Cummins, R. (2010) 'The world in the head' Oxford University Press

Davidson, D. (1963) 'Actions, Reasons, and Causes' in Davidson, D. (2006) 'The Essential Davidson' Oxford University Press

Davidson, D. (1974) 'Thought and Talk' in Davidson, D. (2001) 'Inquiries into truth and interpretation' Oxford University Press

De Houwer, J. (2014) 'A Propositional Model of Implicit Evaluation' *Social and Personality Psychology Compass,* vol. 8 issue 7 p. 342-353

Dennett, D. C. (1978) 'Beliefs about beliefs' *Behavioral and Brain Science,* ***4****,* p. 568-570

Dennett, D. C. (1991) 'Two Contrasts: Folk Craft Versus Folk Science, and Belief versus Opinion.' In Greenwood, J. D. (ed.) (1991) 'The Future of Folk Psychology: Intentionality and Cognitive Science, p135-48. Cambridge University Press, 1991

Evans, J. St. B T & Frankish, K. eds. (2009) 'In Two Minds: Dual Processes and Beyond' Oxford University Press, Oxford.

Flavell, J. H., Green, F. L & Flavell, E. (1990) 'Developmental Changes in Young Childrens Knowledge About the Mind' *Cognitive Development*, **5**, p. 1-27

Fletcher, L. & Carruthers, P. (2013) 'Behaviour Reading versus Mentalizing in Animals' in Metcalfe, J. & Terrence, M. (Eds.), 'Agency and Joint Attention' Oxford University Press, Oxford

Fodor, J. (1974) ‘Special sciences (or: the disunity of science as a working hypothesis)’ *Synthese*, vol. 28, no. 2 p. 97-115

Fodor, J. (1983) ‘The Modularity of Mind’ MIT Press, Cambridge Massachusetts

Fodor, J. (1987) ‘Psychosemantics: The Problem of Meaning in the Philosophy of Mind’ MIT Press, Cambridge Massachusetts.

Fodor, J. (1997) ‘Special sciences: still autonomous after all these years’ *Philosophical Perspectives*, vol. 11, p. 149-163

Frankish, F (2004) ‘Mind and Supermind’, Cambridge University Press, Cambridge.

Ferguson, M. J., Mann, T. C., Cone, J. & Shen, X. (2019) ‘When and How Implicit First Impressions Can Be Updated’ *Current Directions in Psychological Science*, vol. 28 (4), p. 331-336

Gendler, T. (2008a) ‘Alief in Action (and Reaction)’ *Mind and Language*, **23**, p. 552-585

Gendler. T. (2008b) ‘Alief and Belief’ *The Journal of Philosophy*, **105**, p. 634-663

Gergely, G. Nadasdy, Z. Csibra, G & Biro, Z. (1995) ‘Taking the intentional stance at 12 months of age’ *Cognition*, **56**, p. 165-193

Gettier, E. (1963) ‘Is Justified True Belief Knowledge?’ in Sosa, E., Kim, J., Fantl, J., and McGrath, M. (eds.) (2008) ‘Epistemology: An Anthology’ Blackwell Publishing. p. 192-193

Gopnik, A. & Astington, J. W. (1988) ‘Children’s understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction.’ C*hild Development*, **59**, p. 26-37

Gopnik, A. (1998) ‘Explanation as orgasm’ *Minds and Machines*, **8**, p. 101-118

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L. and Banaji, M. R. (2009) ‘Understanding and using the implicit association test, III: meta-analysis of predictive validity’, *Journal of Personality and Social Psychology*, **97**, p. 17-41

Grice, H. P. (1975) ‘Logic and conversation’ in Grice, H. P. (1989) ‘Studies in the way of words’, Harvard University Press

Griffiths, P. (1997) *What emotions actually are*, University of Chicago Press, Chicago

Griffiths, P. (2002) ‘What is innateness?’ *Monist*, vol 85 no. 1 p. 70-85

Griffiths, P., Machery, E. & Linquist, S. (2009) ‘The Vernacular Concept of Innateness’ *Mind and Language*, Vol. 24 issue 5, p. 605-630

Hannan, B. (1993) ‘Don’t stop believing: the case against eliminative materialism’ *Mind and Language*, Vol. 8 issue 2, p. 165-179

Hare, B., Call, J., Agnetta, B. & Tomasello, M (2000) ‘Do chimpanzees know what conspecifics know?’ *Animal Behaviour*, **59**, p. 771-785

Hare, B., Call, J. & Tomasello, M. (2001) ‘Do chimpanzees know what conspecifics know?’ *Animal Behaviour*, **61**, p. 139-151

Harman, G. (1965) ‘The Inference to Best Explanation’ *The Philosophical Review*, vol. 74, no. 1, p. 88-95

Harman, G. (1978) 'Studying the champanzee's' theory of mind'. *Behavioral and Brain Science,* ***4****,* p. 568-570

Helming, K. A., Strickland, B., & Jacob, P. (2014) 'Making sense of early false-belief understanding', *Trends in Cognitive Sciences, vol. 18, no. 4,* p. 167-170

Hogrefe, G. J., Wimmer, H. & Perner, J. (1986) 'Ignorance versus false belief: A developmental lag in attribution of epistemic states' *Child Development*, **57**, p. 567-582.

Hudson, R. (2020) 'The Reality of Jean Perrin's Atoms and Molecules' *The British Journal for the Philosophy of Science,* vol. 71, **1**, p. 35-58

Hyde, D. C. & Spelke, E. S. (2011) 'Neural signatures of number processing in human infants: evidence for two core systems underlying numerical cognition' *Developmental Science*, vol. 14 (2) p. 360-371

Ichikawa, J. J. & Steup, M. (2018) 'The Analysis of Knowledge', in Zalta, N. (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition) URL = <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>.

International Astronomical Union (2006) 'Pluto and the Developing Landscape of Our Solar System' view 10/12/20 https://www.iau.org/public/themes/pluto/

Jenson, J.C. (2016) 'The Belief Illusion' *British Journal for the Philosophy of Science*, **67**, p. 965-995

Jones, T. E. (2004) 'Special Sciences: still a flawed argument after all these years' *Cognitive Science*, **28**, p. 409-432

Király, I., Oláh, K., & Kovács, A. M. (2018) 'Retrospective attribution of false beliefs in 3-year-old children' *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 45, p. 11477-11482

Kovacs, A. M., Teglas, E. & Endress, A. D. (2010) 'The Social Sense: Susceptibility to Others' Beliefs in Human Infants and Adults' *Science*, vol. 330, p. 1830-1834

Kuhn, T (1996) [1962] 'The Structure of Scientific Revolutions', The University of Chicago Press, Chicago.

Lewis, D. (1982) 'Logic for equivocators' *Nous*, vol. 16 (3), p. 431-441

Lillard, A. (1998) 'Ethnopsychologies: cultural variations in theories of mind', *Psychological Bulletin*, vol. 123, No.1, p. 3-32

Mameli, M. & Bateson, P. (2006) 'Innateness and the sciences' *Biology and Philosophy*, **21**, p. 155-188

Madelbaum, E. (2016) 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias' *Nous*, 50:3, p. 629-658

Melis, A., Call, J. & Tomasello, M. (2006) 'Chimpanzees (*Pan troglodytes*) conceal visual and auditory information from others' *Journal of Comparative Psychology*, **120**, p. 154-162

Moll, H., Khalulyan, A., & Moffat, L. (2017) '2.5-Year-Olds Express Suspense When Others Approach Reality With False Expectations' *Child Development*, vol. 88, no. 1, p. 114-122

Myers-Schulz, B. & Schwitzgebel, E. (2013) 'Knowing That P without Believing That P', *NOUS, 47*:2 p. 371–384

Nemeroff, C. & Rozin, P. (1994) 'The contagion concept in adult thinking in the United States: transmission of germs and interpersonal influence' *Ethos*, **22**, p. 158-86

Nichols, S. & Stich, S. P. (2003) ‘Mindreading: An integrated account of pretence, self-awareness and understanding other minds’ Oxford University Press, Oxford.

Nisbett, R. & Wilson, T. (1977) ‘Telling more than we can know: Verbal reports on Mental Processes’ *Psychological Review*, **84**, p. 231-259

Onishi, K. H. & Baillargeon, R (2005) ‘Do 15-Month-Old Infants Understand False Beliefs?’ *Science*, **308**, p. 255-258

Orwell, G. (1987) [1949] ‘Nineteen Eighty-Four’ Penguin

Perner, J. (1991) ‘Understanding the representational mind’. Cambridge, MA: MIT Press

Perner, J. (2010) ‘Who took the cog out of cognitive science? Mentalism in an era of anti-cognitivism.’ In Frensch, P, & Schwarzer, R. (2010) ‘Cognition and Neuropsychology: International Perspectives on Psychological Science: Volume 1.’ New York, Psychology Press.

Perner, J. & Roessler, J. (2012) ‘From infants’ to children’s appreciation of belief’ *TRENDS in Cognitive Science*, **16**, No. 10, p. 519-525

Perner, J., Leekham, S. & Wimmer, H. (1987) ‘Three-year-olds difficulty with false belief: the case for conceptual deficit’ *British Journal of Developmental Psychology*, **5**, p. 125-135

Perner, J & Ruffman, T. (2005) ‘Infant’s insight into the mind: How deep?’ *Science*, **308**, p. 214-216

Povinelli, D. J. & Vonk, J. (2004) ‘We don’t need a microscope to explore the chimpanzee’s mind’. *Mind & Language*, **19**, p. 1-28

Quilty-Dunn, J. & Mandelbaum, E. (2018) ‘Against dispositionalism: belief in cognitive science’ *Philosophical Studies*, 175, p. 2353-2372

Ravenscroft, I. ‘Folk Psychology as a Theory’, *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition),* Edward N. Zalta (ed.), URL =
<https://plato.stanford.edu/archives/sum2019/entries/folkpsych-theory/>

Roit, I., Brostoff, J. & Male, D. (2001) ‘Immunology (sixth edition)’ Mosby

Ross, L., Lepper, M.R. and Hubbard, M. (1975) ‘Perseverance in Self-perception and Social Percption: Biased Attributional Processes in the Debriefing Paradigm’ *Journal of Personality and Social Psychology*, **32**, p. 880-892

Rozin, P., Millman, L. & Nemeroff, C. (1986) ‘Operation of the laws of sympathetic magic in disgust and other domains’ *Journal of Personality and Social Psychology*, **50**, p. 703-712

Ruffman, T. & Perner, J. (2005) ‘Do infants really understand false belief?’ *TRENDS in Cognitive Science*, vol. 9, no. 10, p. 462-463

Ryle, G. (1963) [1949] ‘The Concept of Mind’ Penguin

Samuels, R. (2007) ‘Is Innateness a Confused Concept?’ in Carruthers, P., Laurence, S. & Stich, S. (eds.) (2007) ‘The Innate Mind: Foundations and the future’ Oxford University Press, p. 17-36

Saxe, R (2005) ‘Against Simulation: The Argument from Error’ *TRENDS in Cognitive Sciences*, Vol. 9, No. 4, p. 174-178

Schneider, D., A. Bayliss, S. Becker, & P. Dux. (2012) ‘Eye movements reveal sustained implicit processing of others’ mental states’ *Journal of Experimental Psychology: General* vol. 141(3), p. 433–438.

Schwitzgebel, E. (2002) ‘A Phenomenal, Dispositional Account of Belief’ *Nous*, 36:2, p. 249-275

Schwitzgebel, E. (2003) ‘In-between believing’ *The Philosophical Quarterly*, vol. 51, no. 202, p. 76-82

Schwitzgebel, E. (2010) ‘Acting contrary to our professed beliefs or the gulf between occurrent judgement and dispositional belief’ *Pacific Philosophical Quarterly*, **91**, p. 531-555

Schwitzgebel, E. (2013) ‘A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box’ in Nottleman, N. (ed.) (2013) ‘New Essays on Belief’ Palgrave MacMillan, p. 75-99

Schwitzgebel, Eric, (2019) ‘Belief’, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/belief/>. [Jenson references the 2011 version of this article – the account of the functional role of belief is identical in both versions]

Scott, R. M. & Baillargeon, R. (2009) ‘Which Penguin Is This? Attributing False Beliefs about Object Identity at 18 Months’ *Child Development*, **80**, No.4, p. 1172-1196

Scott, R. M. & Baillargeon, R. (2017) ‘Early false-belief understanding’ *Trends in Cognitive Sciences*, vol. 21, no. 4, p. 237-249

Scott, R. M., Baillargeon, R., Song, H. & Leslie A. M. (2010) ‘Attributing false beliefs about non-obvious properties at 18 months’ *Cognitive Psychology*, **61**, p. 366-395

Sellars, W. (1997) [1956] ‘Empiricism and the Philosophy of Mind’ Harvard University Press

Setoe, P., Scott, R. M. & Baillargeon, R. (2016) ‘Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands’ *Proceedings of the National Academy of Sciences of the United States of America*, vol.113 (47), p. 13360-13365

Song, H., Onishi, K. H., Baillargeon, R. & Fisher, C (2008) ‘Can an agent’s false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants’ *Cognition*, **109**, p. 295-315.

Song, H. & Baillargeon, R. (2008) ‘Infants' reasoning about others' false perceptions’ *Developmental Psychology,* 44 (6), p. 1789–1795.

Southgate, V. (2013) ‘Early manifestations of mindreading’ in Baron-Cohen, S., Tager-Flusberg, H., and Lombardo, M. V. (eds.) (2013) ‘Understanding other minds: perspectives from developmental social neuroscience – third edition’ Oxford University Press, p. 3-18

Southgate, V. Senju, A. & Csibra, G. (2007) ‘Action Anticipation Through Attribution of False Belief by 2-Year-Olds’ *Psychological Science,* **18**, No.7, p. 587-592

Southgate, V. Chevallier, C. & Csibra, G. (2010) ‘Seventeen-month-olds appeal to false beliefs to interpret others’ referential communication’ *Developmental Science*, **13**, p. 907-91

Southgate, V. & Vernetti, A. (2014) ‘Belief-based action prediction in preverbal infants’ *Cognition*, **130**, p. 1-10

Stammers, S. (2017) ‘A Patchier Picture Still: Biases, Beliefs and Overlap on the Inferential Continuum’ *Philosophia*, vol. 45, p. 1829-1850

Sterelny, K. (2003) ‘Thought in a Hostile World: The Evolution of Human Cognition’ Blackwell Publishing

Stich, S. P. (1978) ‘Beliefs and subdoxastic states’ *Philosophy of Science*, vol. 45 no. 4, p. 499-518

Stich, S. P. (1996) ‘Deconstructing the Mind’ Oxford University Press, Oxford

Tomasello, M., Call, J. & Hare, B (2003) ‘Chimpanzees understand psychological states – the question is which ones and to what extent’ *TRENDS in Cognitive Sciences*, **7**, No.4, p. 153-156

Tversky, A., Sattath, S. & Slovic, P (1988) ‘Contingent Weighting in Judgment and Choice’ Psychological Review, vol. 95 (3), p. 371–384

Tversky, A. & Thaler, R. H. (1990) ‘Anomalies: preference reversals’ *Journal of Economic Perspectives* vol. 4, no. 2, p. 201-211

Van Dessel, P., De Houwer, J., Gast, A. & Tucker Smith, C. (2015) ‘Changing Stimulus Evaluation via the Mere Instruction to Approach or Avoid Stimuli’ *Experimental Psychology* vol. 62 (3), p. 161-169

Van Dessel, P., De Houwer, J., Gast, A., Tucker Smith, C. & De Schryver, M. (2016) ‘Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation’ *Journal of Experimental Social Psychology*, vol. 63, p. 1-9

Van Dessel, P., Ye, Y. & De Houwer, J. (2019) ‘Changing Deep-Rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi’ *Social Psychological and Personality Science*, vol. 10(2), p. 266-273

Vinden, P. G. (1996) ‘Junin Quechua Children’s Understanding of Mind’ *Child Development*, vol. 67 no. 4, p. 1707-1716

Wellman, H. M., Cross, D. & Watson, J. (2001) ‘Meta-analysis of theory of mind development: the truth about false belief’. *Child Development*, **72**, p. 655-684

Wellman, H. M. (2014) ‘Making minds: how theory of mind develops’ Oxford University Press

Whiten, A (1996) ‘When does smart behaviour reading become mind reading?’ in Carruthers, P & Smith, P.K. (eds.) (1996) ’Theories of Theories of Mind’ Cambridge University Press, p. 277-292

Wilkes, K. (1993) ‘The relationship between scientific and common sense psychology’ in Christensen, S. & Turner, D. (eds.) (2003) ‘Folk psychology and the philosophy of mind’ Hillsdale: Lawrence Earlbaum, p. 144-187

Wilby, M (2012) ‘Embodying the False-Belief Tasks’ *Phenomenology and the Cognitive Sciences*, **11** (4), p. 519-540

Wilson, T. (2002) ‘Strangers to Ourselves: Discovering the Adaptive Unconscious’ Belknap, Harvard

Wimmer, H. & Perner, J. (1983) ‘Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception’ *Cognition*, **13**, p. 103-128

Wimsatt, W.C. (1982) ‘Robustness, Reliability, and Overdetermination’ in Brewer, M. & Collins, B. (eds) *Scientific Inquiry and the Social Sciences*, San Francisco: Jossey Bass, p. 124-163

Wittgenstein, L. (2009) [1952] ‘Philosophical Investigations’. Wiley-Blackwell

Woodward, J. (2006) ‘Some varieties of robustness’ *Journal of Economic Methodology*, 13:2, p.219-24

Zawidzki, T. W. (2013) ‘Mindshaping’ MIT Press

Acknowledgements

Huge thanks to my supremely patient supervisor George. I’ve never come out of a meeting with him feeling less positive than I went in. Also, many thanks to Stephen and Jerry, who provided support and a different perspective. Finally, to Clare, for her support and encouragement.